Middle-East Journal of Scientific Research 19 (2): 258-267, 2014
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2014.19.2.11919

Achievement vs Proficiency: Construction of Math Proficiency Assessment Using Item Response Theory

[1]Muhammad Azeem, [2]Muhammad Bashir Gondal and [1]Abdullah Faisal

[1]Punjab Education Assessment System [PEAS], College Road Township, Lahore, Pakistan
[2]Punjab Education Commission, Wahdat Colony Road, Lahore, Pakistan

Abstract: Assessment is a systematic process that begins with identification of objectives and ends with a judgment concerning the extent to which those objectives have been attained. Achievement assessment means assessment of what has been taught or studied within the modules. Proficiency assessment is an exercise the purpose of which is to evaluate the background of a person in a given branch of knowledge with or without regard to specific academic. There is distinction between proficiency test and achievement test. One measures a student's ability to use something and the other measures a student's knowledge or at least recall. The most important reason to select IRT is that IRT can be gainfully applied: estimating examinee ability, generalizability of test result, various item analyses, test and item scaling, and it provides a way of giving meaningful score interpretations relating students' performance to the underlying skills being measured. Total 2680 students were included in the sample for study using proportionate stratified cluster random sampling technique. National Assessment of Educational Progress (NAEP)'s mathematics assessment framework is adapted as math proficiency framework based on objectives of Pakistan's national math curriculum for 9th grade. Valid and reliable math proficiency consisted of 60 items aligned with content and performance standards of mathematics for 9th grade and adapted mathematics framework was developed after tryout and piloting of 190 items.

Key words: Item Response Theory IRT · Measurement · Psychometric · Testing · Assessment · Test Score

INTRODUCTION

Assessment is a systematic process that plays a significant role in effective teaching. It begins with identification of objectives and ends with a judgment concerning the extent to which those objectives have been attained (PEAS Report, 2005). In simplest words, assessment provides information on whether teaching/ learning has been successful. However, the information it provides has a number of potential different audience whose precise requirements may vary [1]. The main purpose of development of an assessment is to set priorities and guide about whole scenario of the context of use of assessment design. Thus, an assessment is purpose-specific and context-specific although it may use for other contexts e.g. proficiency assessment may used to measure achievement or a single assessment is used for individual student and programme evaluation. Assessment, for whatever purpose, is a process of reasoning from evidence [2] Glaser, 2001). Assessment fulfills the "desire to reason from particular things students say, do, or make, to inferences about what they know or can do more broadly" [3].

Brindley (1998), as cited in [4-7], reports that the "distinction between assessing proficiency and assessing achievement is becoming increasingly blurred in a number of ways. Some methods used to assess ongoing achievement are virtually indistinguishable from those used to assess proficiency. This applies particularly to criterion-referenced forms of assessment in which domains of language ability and standards of performance are well-defined (p.3)". Therefore, assessment of proficiency and assessment of achievement may differentiate.

Assessment of Achievement: According to Sestic and Huttunen (2006) [8] achievement assessment means assessment of what has been taught/studied within the

Corresponding Author: Muhammad Azeem, Punjab Education Assessment System [PEAS], College Road Township, Lahore, Pakistan.

modules; it is oriented internally i.e. to classroom activities, the course book, the syllabus. It refers to the achievement of set objectives and is usually under taken within a particular class [9]. Argue that achievement tests are directly related to the courses, their purpose being to establish how successful individual students, groups, or the courses themselves have been in achieving objectives [10]. Describes that achievement tests directly related to courses, their purpose being to measure the extent of learning in a prescribed content domain, often in accordance with explicitly stated objectives of a learning programme. McNamara (2002) is of the view that achievement tests are associated with process of instruction….support to teaching to which they relate…reflect progressive aspects of curriculum…relate to the past in that they measure what the students have learned as a result of teaching.

Thus, the achievement tests test the material taught. In fact, it measures the students' learning process. Achievement assessment looks at what students know rather than at what they can do. It requires students to demonstrate retention of previously learned content material.

Assessment of Proficiency: Proficiency assessment is an exercise the purpose of which is to evaluate the background of a person in a given branch of knowledge with or without regard to specific academic learning [11]. According to [8] proficiency assessment is oriented externally; independent of the syllabus; related to proficiency level. Its intention is not to put learners into rank order but to assess what student can do, how much competence he can put to use, how well he applies his knowledge in the real world. Proficiency testing is standard based assessment. The required standards assessed through proficiency testing. Proficiency describes a discipline-related skill, competency or ability that in ordinary environment, a learner should have in order to expect successful participation in the next level or area of study. Proficiency tests need standards and benchmarks, which guide the development of test blueprints, task specifications and proficiency measures. The emphasis in proficiency assessment is on performance.

Distinction Between Assessment of Proficiency and Achievement: There is distinction between proficiency test and achievement test. One measures a student's ability to use something and the other measures a student's knowledge or at least recall. Proficiency tests are designed to measure students' achievements in relation to a specific task, which they are later required to perform while achievement tests are designed to assess what a student has learned, or at least what a student knows [11]. [12] proficiency tests look to the future situation without necessarily any reference to the previous process of teaching…include performance feature in their design—for example a test of communicative abilities of health professionals in work setting will be based on representations of such workplace tasks as communicating with patients or other health professionals...have gate keeping function-for example admission to an overseas university, or to an occupation requiring practical second language skills.

As proficiency, assessment is standards based assessment and globally it is being accepting as assessment tool that support learning and instruction therefore it our need to evaluate our students using proficiency assessments.

Assessment of Proficiency and IRT: The achievement of students can be assessed by classical test theories (CTT) in which all easy and difficult items have same weightage for all students. CTT does not take into account the difficulty levels of the items when estimating learning achievement of students. A student scoring 10 for 10 easy items does not imply equal learning achievement/proficiency to another scoring 10 for 10 difficult items. The focus of CTT is most often on single score that one obtains on a test. It treats all items as though they were parallel [13]. Logically, getting a more difficult item correct should be given a 'higher' weighting when measuring learning achievement/proficiency. The focus of IRT is on the pattern of responses that the respondent makes to the set of items and it does not assume that all items on the test are parallel [13]. Thus, CTT and IRT provide different and complementary ways to examine the responses to a series of items. The person's ability and item difficulty locating on the same continuum of measure is the essence of IRT. IRT helps in estimating abilities of examinees based on item difficulty, item discrimination and the guessing or chance factor as well. Rash model is most popular IRT model used for data analysis. Rash model take in to account only items' difficulty in estimating students' proficiency. It measures students' ability and items' difficulty on same scale.

IRT measures the probable ability of student according to the responses against items of predetermined difficulty level. According to Carroll (1996),

ability is the capability to perform some task, where there are "possible variations over individuals in their threshold levels of difficulty in successfully performing some defined class of tasks. Through IRT, task demands can be directly linked to ability scores so that meaningful feedback can be given to students.

The most important reason to select IRT is that IRT can be gainfully applied: estimating examinee ability, generalizability of test result, various item analyses, test and item scaling and it provides a way of giving meaningful score interpretations relating students' performance to the underlying skills being measured. This is particularly helpful if the purpose of an assessment is to inform teachers and students of learning strategies. IRT models, estimate students' abilities and item difficulties in a single mathematical model, allowing for a proficiency continuum to be developed, with students' abilities and item difficulties placed on the same scale. Thus once a students' ability is located on the scale, it is possible to make inferences about the kinds of tasks the student is likely to be able to perform. In the most general sense, IRT is a psychologically based theory of mental measurement that specifies information about latent traits and characteristics of stimuli (test items) used to represent them. It maintains that there is an estimable latency for an examinee to any particular proficiency that covers the range (-8, 8) and that its estimation does not depend on particular items or assessment exercises [13] [14] have published a problem-solving assessment kit for which IRT modeling used to develop a proficiency scale.

Assessment of Math Achievement: Making sure the development of mathematical competencies during the primary grades is essential to later learning success. Most of the children who have had less experience or exposure to mathematical concepts fail to meet minimal mathematics proficiency standards by the end of their formal schooling and are at high risk for mathematics failure. It is because math is highly procedural and continually builds on previous knowledge for successful learning therefore early deficits have enduring and devastating effects on later learning. Early mathematics interference and diagnoses can repair deficits and prevent future deficits. Deficits can identify by valid assessment technique. Assessment is a way of measuring what students know and of expressing what students should learn. Mathematics education has changed as the role of mathematics in society has changed. In Mathematics education, curriculum and instruction is based on standards. Therefore, Mathematics assessment has to ensure consistency with the goals of education. Students think well when they learn better and they learn better when they utilize their minds and think. Assessment has pivotal role in education just as standards provide guidance for reform. Basically, assessment tells stakeholders about students learning. Assessment should do much more than test discrete procedural skills. Assessment should see as an integral part of learning rather than a separate tool. The goal of mathematics learning cannot be realized if assessment remains to measure that is easy and has traditionally been taught. Therefore, math assessment should emphasize problem solving, thinking and reasoning. Assessment of mathematics should require students to communicate clearly their mathematical thinking for non-routine problems in real life. It must enable students to construct new knowledge from what they know.

Background of Study: Like many developing countries, Pakistan is also facing a problem of expanding enrollment while at the same time improving the quality of education remains a challenge. Little evidence however, is available in Pakistan on the assessment of quality of student learning. To fill-in this gap, a well-planned and properly executed national assessment is needed. Considering this need, the government of Pakistan is committed to improve the quality of education along with the effort to increase enrollment and access. This commitment is reflected in the government's policy documents (National Educational Policy 1998-2010, Education Sector Reforms 2001-2005) and in its reaffirmation at international forums (Jomtien Declaration 1990 and EFA Assessment 2000, Dakar) to include quality outcomes such as student achievement scores as performance monitoring indicators in the education sector. [16] Although Government of the Punjab has provided teacher training in the subject of science and mathematics up to secondary school level to ensure the quality education in the subject of science and math, our existing assessment system still encourages memorization and mechanical reproduction of texts. It only provides little information and does not depict the competencies and skills that have been achieved by the students at a specific level. It also does not help in improving the quality of education at large scale. It is the need of day, to assess our students as a product not on the basis of their learning process but on the bases how they emphasize on problem solving and how they are thinking critically with reasoning. Assessment tools should be developed to measure what they can to do (perform) rather than what they know to do. How they

communicate clearly their mathematical thinking for non-routine problems in real life. Since proficiency is person-specific and context-specific [17] and proficiency test measures students' achievement in relation to a specific task which they are later required to perform (Power, n.d) therefore it is time to shift our assessment system from assessing the learning of the students towards assessing the proficiencies of the students. An assessment system that can assess students' proficiency is necessary to compare our students' academic performance with internationally accepted proficiency levels.

Internationally, different countries are shifting their assessment system from traditional achievement testing towards proficiency testing. Australia, USA, Canada and European countries are at leading position in proficiency testing. Research at organizational level is being carried out in field of proficiency assessment. Emphasis is especially on English as a Second Language (ESL), Science and Mathematics. Mathematics proficiencies are being assessed to know how students are performing in the crucial subject of mathematics.

Unfortunately, in Pakistan, students are being assessed by traditional methods of achievement testing. Up to grade 12, the students are taking Boards Exams. Boards, due to lack of research in educational assessment, are using non-standardized contents based teacher made achievement tests. These tests are assessing students' retention power rather than their performance. Furthermore, there may be any example of proficiency testing or study is in Pakistan. There is a gap between notion of assessing students at nationally and internationally. Therefore, to fulfill this gap, it is our need to adopt the notion of proficiency testing so that we may be able to assess and compare our students at national and international levels and to help our students and teachers in improving quality of education.

MATERIALS AND METHODS

Sampling: This study is related to research and development in the field of psychometric. All the 9th grade students studying in public high and higher secondary schools in the province of Punjab were comprised the population of the study. Locale and gender were the strata of sampling. Total 2680 students were included in the sample for study. The sample was chosen by considering minimum sample size for each stratum as basic requirement of Item Response Theory based software ConQuest. For this study more than 500 students are required from each four strata for IRT based analysis. According to Sample Design Table (Appendix A) with ρ (roh) =0.3 and cluster size 20 from 134 schools the sample size is 2680. This 2680 sample size was selected using proportionate stratified cluster random sampling technique.

Instrumentation: National Assessment of Educational Progress (NAEP)'s mathematics assessment framework is adapted for this study. NAEP's mathematics assessment framework is modified as math proficiency framework (Appendix B) based on objectives of Pakistan's national math curriculum for 9th grade. Total 196 Test items were developed according to the dimensions of developed math proficiency framework. 190 test items were aligned with math proficiency framework and national math curriculum. By consensus method, 10 items were selected as common items for three test booklets. Remaining one hundred and eighty items were used to develop three parallel math proficiency tests. Remaining items were sorted out in five clusters on the basis of strands. 60 items were selected among the clusters randomly for each parallel test. The items to these three parallel tests were assigned randomly. These three parallel tests were divided in to equal halve blocks. Different two halves were recombined to develop final three booklets. Common ten items were added to each booklet. Final three booklets containing 70 items with 10 common items were developed for piloting. The data of pilot study was analyzed by ITMAN and ConQuest software. Difficulty and discrimination indices of each item were analyzed. By setting difficulty rang 0.25 to 0.8, taking Item difficulty (b) value in range -3.00 to +3.00, Infit and outfit mean square values range 0.80 to 1.2 and expert opinion, 60 items were validated and selected. So the final Math Proficiency test was consisted of 60 items.

Finalization of Items for Final Test: Outputs from IRT based software ConQuest and classical theory based software ITMAN were analyzed. Difficulty levels of all items corresponding ability levels of students were analyzed. Distractors analyses were carried out.

Key indices used for finalization of items for final data are given in Table 1.

Items were finalize those fulfill all six conditions with maximum priority to those fulfill at least IRT based conditions with minimum priority along with Webb's alignment criteria. 60 items for the final test were selected. All 10 common items and 50 out of 180 items were selected for final test.

Table 1: Item Selection Criteria

IRT Based	CTT Based
Range of "b" -3 to +3	Difficulties Index 0.30 to 0.80
Mean-Square Range of "Infits" 0.80 to 1.20	Discrimination 0.20 to 0.80
Mean-Square Range of "Outfits" 0.80 to 1.20	Point-biserial > 0.8

Table 2: Alignment Evidence

		Number of items and associative Proficiency				
Content Strands	Benchmarks/ Indicators	Problem Solving	Procedural Knowledge	Conceptual Understanding	Proficiency Total	Strand Total
Numbers & Sets	SN1		1	1	2	16
	SN2		1	1	2	CU 6
	SN3		1		1	PK 8
	SN4		1	1	2	PS 2
	SN5		1	1	2	
	SN6	1	1	1	3	
	SN7		1	1	2	
	SN8	1	1		2	
Algebra	A1		1	1	2	12
	A2	1	1		2	CU 5
	A3	1		1	2	PK 4
	A4		1	1	2	PS 3
	A5		1	1	2	
	A6	1		1	2	
Matrices & Logarithm	ML1		1	1	2	20
	ML2		1	1	2	CU 9
	ML3	1		2	3	PK 7
	ML4		1	1	2	PS 4
	ML5		1	1	2	
	ML6	1	1		2	
	ML7	1		1	2	
	ML8		1	1	2	
	ML9		1	1	2	
	ML10	1			1	
Geometry & Measurement	GM1			2	2	12
	GM2		1	1	2	CU 6
	GM3		1	1	2	PK 4
	GM4		1	1	2	PS 2
	GM5	1		1	2	
	GM6	1	1		2	
Total		11	23	26	60	

Alignment of Items and Proficiency Framework

Table 3: Reliability Indices Generated by Different Software

Reliability	Index	Software
Item Separation reliability	0.999	Conquest
MLE Person separation Reliability:	0.875	Conquest
WLE Person separation Reliability:	0.872	Conquest
Cronbach Coefficient Alpha	0.890	Conquests
Cronbach Coefficient Alpha	0.881	SPSS
Marginal Reliability (1PL)	0.8841	Multilog
Marginal Reliability (2PL)	0.9046	Multilog
Marginal Reliability (3PL)	0.9990	Multilog

Content and Skill Areas: The final test covers the following contents and skills as shows in Table 2.

Administration of Math Proficiency Test: Math proficiency test was administered in 2680 students of 134 schools of Punjab province. 134 test administrators were worked as voluntary data collectors. Data collectors were trained for test administration, coding and marking of the test. Data was analyzed by using SPSS 12, ConQuest and Multilog software. Only 3.62% was missing. The missing data was not considered in the analyses.

Data Analyses: Overall reliabilities (Table 3) of test items were reasonably high for 60-items math proficiency test

suggesting that there is internal consistency across items. Item and person separation reliabilities suggested that overall test is reliable and valid regardless of the model fit.

Both infit and outfit statistics indices provide evidence of goodness-of-fit. The reliability and goodness-of-fit of data collected by administering math proficiency test is checked by mean squares fit statistics and reliability indices. For 60-items math proficiency test 59 items were within the mean squares fit statistics (infit) range ≥0.8 to ≤1.3 while only one item was of 1.35 infit mean square values. Thirty six items were below 1.00, three were at 1.00 and remaining twenty one items were above 1.00. Majority of the items were within the range 1±0.05 i.e. within 5% confidence interval. For outfit statistics 54 items were within rang ≥0.8 to ≤1.3, six items have outfit values ≥1.3. 25 items were above 1.00, two items were at 1.00, while remaining 23 items were below 1.00. Therefore, all assumptions of Rash model were met well. Items' difficulties were also within ±3. Most of the items (50 items) have discrimination power ≥0.3 to ≤0.5 while remaining 10 items were below 0.3 discrimation power value including two items with negative value. These infit and outfit statistics along with items' difficulty and discrimination power range suggested the validity and reliability of t test items. The mean scaled score of overall math proficiency test, Multiple Choice MC items and Short Constructed Response SCR items was 499, 476 and 579 respectively. It may explore that MC items are relatively hard than SCR items or students were more familiar with SCR items than MC items.

Overall Test: Item-person map (Figure 1) generated by ConQuest software explores that all items were within difficulty range ±3 with 33 items out of 60 items were within difficulty range ±2. It shows nearly normal distributions of items and students on proficiency scale which ensure the reliability of the test. It also explores that most of the problem-solving items were above the mean of the ability scale that is items for higher ability levels were located on ability scale at higher order. Items measuring students' conceptual understanding and procedural knowledge show mixed results. It is a common thinking that items measuring students' procedural knowledge are of moderate difficult than items measuring students' conceptual understanding. This map explores that some items measuring students' conceptual understanding are harder than that items measuring students' procedural knowledge. This test explores that majority of the items (14 out of 26) measuring students' conceptual understanding are harder than majority of the items (14 out of 23) measuring students' procedural knowledge. It may due as indicated in the literature review that some researches proved that conceptual understanding is necessary before performing procedures but others researches argue that procedural knowledge leads to conceptual understanding. Alignment of all the 60 items with math proficiency framework and national math curriculum for secondary classes ensures the validity of the test. Fit statistics of test items shows that all items have goodness of fit with Rash model. Fit statistics also provide the bases for validity of the test. Reliability indices proved the reliability of test.

Final Math Proficiency Test Sun Apr 22 21:59 2009
Map of Latent Distributions and Response Model Parameter Estimates

Terms in the Model (excl Step terms)
Students of 9th class +item

```
| |
| |
3 | |
| |
X|SN7PK07 |
| |
XX| |
| |
XX| |
XXXX| |
XXXX|ML1PK10 |
2 XXXX|GM6PS11 |
XXXX| |
XXXXXX|ML10PS03 SN8PS07 |
```

Each 'X' represents 4.0 cases

Fig. 1: ITEM-PERSON DISTRIBUTION MAP

Hypotheses Testing: In previous discussions in chapter 4 empirically explored that short constructed response items were easier than multiple choice items. Items of procedural knowledge and items of problem solving were reasonably correlated. Items are well aligned with math proficiency framework and math curriculum. These all are evidence of valid and reliable test construction. It may prove the internal validity and reliability of the test. External validity and reliability may be judged by testing hypotheses. All eight hypotheses as stated in chapter I were tested separately. The analysis of each hypothesis is as under:

Ho: There is no significant difference of math proficiency test score among male and female students of 9th grade.

An independent samples *t* test was performed comparing the male students' mean math proficiency test score (M = 504.57, SD = 84.56) with that female students' mean math proficiency test score (M = 503.01, SD = 84.81). The alpha level was .05. This test was found to be statistically not significant, $t(2615) = -0.47$, $p=0.64$ however it indicates that male students show slightly better math proficiency than female students. The same math proficiency test scores of male and female students may indicates that classroom instructional strategies are almost same. It may explore that the math teachers in male and female schools are almost competent in their subject. It can be conclude that math proficiency is independent of gender.

Ho: There is no significant difference of math proficiency test score between rural and urban students of 9th grade.

An independent samples *t* test was performed comparing the urban students' mean math proficiency test score (M = 507.42, SD = 85.25) with that rural students' mean math proficiency test score (M = 499.84, SD = 83.87). The alpha level was .05. This test was found to be statistically significant, $t(2615) = -2.29$, $p=0.022$. This result indicates that math proficiency of urban students is better than rural students. The effect size $r=0.04$ with $d=0.09$ indicates that the distributions of math proficiency scores of both groups overlap completely and means of math proficiency scores of urban students is at 50th percentile of the rural students. It may explore that the math teachers in urban and rural schools have not same level of competency in their subject or there is shortage of math teachers in rural areas. Better salary package and better living facilities of urban areas are the main reasons of shortage of math teachers in rural areas. Students of urban areas are, often, availing extra coaching via tuition centers or their family members therefore they show better results.

Ho: There is no significant difference of math proficiency test score between rural female and rural male students of 9th grade.

An independent samples *t* test was performed comparing the rural male students' mean math proficiency test score (M = 502.55, SD = 70.43) with that rural female students' mean math proficiency test score (M = 497.60, SD = 93.49). The alpha level was .05. This test was found to be statistically not significant, $t(1238) = -1.03$, $p=0.30$. This result indicates that math proficiency of male students is slightly better than female students.

Ho: There is no significant difference of math proficiency test score between urban female and urban male students of 9th grade.

An independent samples *t* test was performed comparing the urban male students' mean math proficiency test score (M = 506.00, SD = 93.24) with that urban female students' mean math proficiency test score (M = 509.36, SD = 72.91). The alpha level was .05. This test was found to be statistically not significant, $t(1375) = 0.72$, $p=0.47$. This result indicates that math proficiency of urban female students is slightly better than urban male students.

Ho: There is no significant difference of math proficiency test score between urban female and rural male students of 9th grade.

An independent samples *t* test was performed comparing the urban female students' mean math proficiency test score (M = 509.36, SD = 72.91) with that rural male students' mean math proficiency test score (M = 502.55, SD = 70.43). The alpha level was .05. This test was found to be statistically not significant, $t(1138) = 1.60$, $p=0.11$. This result indicates that math proficiency of urban female students is slightly better than rural male students.

Ho: There is no significant difference of math proficiency test score between rural female and urban male students of 9th grade. The math proficiency of urban male students is slightly better than rural female students.

An independent samples *t* test was performed comparing the rural female students' mean math proficiency test score (M = 497.60, SD = 93.49) with that urban male students' mean math proficiency test score (M = 506.00, SD = 93.24). The alpha level was 0.05.

This test was found to be statistically not significant, $t(1475) = -1.72$, $p=0.08$. This result indicates that math proficiency of urban male students is slightly better than rural female students.

Ho: There is no significant difference of math proficiency test score between rural female and urban female students of 9th grade. The math proficiency of urban female students is better than rural female students.

An independent samples *t* test was performed comparing the rural female students' mean math proficiency test score (M = 497.60, SD = 93.49) with that urban female students' mean math proficiency test score (M = 509.36, SD = 72.91). The alpha level was 0.05. This test was found to be statistically significant, $t(1258) = -2.46$, $p=0.01$. This result indicates that math proficiency of urban female students is better than rural female students.

It may be due to various factors. Non-availability of math teachers in rural areas may the major factor. Most of the schools in rural areas did not have teachers competent in math. In various schools teachers appointed for arts subjects are teaching mathematics. Their instruction is not well effective due to lack of competencies in mathematics.

Ho: There is no significant difference of math proficiency test score between rural male and urban male students of 9th grade.

An independent samples *t* test was performed comparing the rural male students' mean math proficiency test score (M = 502.55, SD = 70.43) with that urban male students' mean math proficiency test score (M = 506.00, SD = 93.24). The alpha level was .05. This test was found to be statistically not significant, $t(1355) = -0.74$, $p=0.46$. This result indicates that math proficiency of urban male students is slightly better than rural male students.

CONCLUSION

Valid and reliable math proficiency aligned with content and performance standards of mathematics for 9th grade and adapted mathematics framework is developed. This math proficiency test covers all content area and proposed math proficiencies. Overall, this test shows high reliability coefficient =0.90. This suggests internal reliability. Webb's alignment model ensures reliability and structural and constructional validity of the test. Fit statistics also validated its structural and constructional validity. It, adequately, distinguishes the students having basic, proficient and advance, pre-fixed, levels in math. This test may have achieved its objectives. Therefore math proficiency test may use for assessing students' math proficiency for various purposes. Similar studies in other subjects and replication of this study is recommended as future studies may validate its external validity.

REFERENCES

1. Fleming, M., 2007. The Challenge of Assessment within Language(s) of Education. Intergovernmental Conference. Languages of schooling within a European framework for Languages of Education learning, teaching, assessment. Prague 8-10 November 2007. Organized by the Language Policy Division, Council of Europe, Strasbourg in co-operation with the Ministry of Education, Youth and Sports, Czech Republic.
2. Pellegrino, J.W., N. Chudowsky and R. Glaser, (Eds.)., 2001. Knowing what students know. The science and design of educational assessment. Washington, DC: National Academy Press.
3. Mislevy, R.J., R.G. Almond and J.F. Lukas, 2004. A Brief Introduction to Evidence-Centered Design. National Center for Research on Evaluation, Standards and Student Testing (CRESST). Center for the Study of Evaluation (CSE). CSE Report 632. University of California, Los Angeles.
4. MacFarlane, A., 2003. Proficiency Test. Retrieved from www.caslt.org/pdf/testproposal.pdf on 15-06-2008.
5. Muhammad Azam, Sallahuddin Hassan and Khairuzzaman, 2013. Corruption, Workers Remittances, Fdi and Economic Growth in Five South and South East Asian Countries. A Panel Data Approach Middle-East Journal of Scientific Research, 15(2): 184-190.
6. Sibghatullah Nasir, 2013. Microfinance in India Contemporary Issues and Challenges. Middle-East Journal of Scientific Research, 15(2): 191-199.
7. Mueen Uddin, Asadullah Shah, Raed Alsaqour and Jamshed Memon, 2013. Measuring Efficiency of Tier Level Data Centers to Implement Green Energy Efficient Data Centers, Middle-East Journal of Scientific Research, 15(2): 200-207.
8. Sestic, L. and I. Huttunen, 2006. A Guide Implementing the Revised Core Curriculum for Modern Languages in Bosinia and Herzegovina. DGIV/EDU/LANG, 2006. 1. Council of Europe.

9. Hughes, A., 2003. Chapter 3. kinds of tests and testing. In Testing for Language Teachers. (2nd ed.) Cambridge University Press, pp: 11-23.
10. Mousavi, S., 1999. A dictionary of language testing. (2nd ed.). Tehran Rahnama Publications.
11. Proficiency and achievement test. Retrieved on 4-16-2006 from https://ask.nsdl.org/default.asps?id=14231&cat=1159.
12. McNamara, T., 2002. Language testing. Oxford, UK: Oxford University Press ISBN: 0194372227.
13. Kline, T.J.B., 2005. Psychological Testing. A practical Approach to Design and Evaluation. New Delhi Sage Publications, Inc.,
14. Osterlind, S.J., 2006. Modern Measurement. Theory, Principal and Applications of Mental Appraisal. New Jersey Person Education, Inc.
15. Stacey, K., S. Groves, S. Bourke and B. Doig, 1993. Profiles of problem solving. Melbourne. Australian Council for Educational Research.
16. National Education Assessment System, 2005. National Assessment Report. Islamabad Ministry of Education, Government of Pakistan.
17. Chastain, K., 1989. The ACTFL Proficiency Guidelines. A selected Sample of Opinions. ADFL Bulletin,.Retrieved April 16, 2006, from http://www.mla.org/adfl/bulletin/v20n2/202047.htm., 20(2): 47-51.
18. Council of Chief State School Officers, CCSSO. September, 2002. Models for alignment analysis and assistance to states. Washington, DC: Author. Retrieved from http://www.ccsso.org/content/pdfs/AlignmentModels.pdf on September 24, 2004.
19. Embretson, S.E., 1991. A multidimensional latent trait model for measuring learning and change. Psychometrika, 56: 495-515.
20. Embretson, S.E., 1995. A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. Journal of Educational Measurement, 32(3): 277-294.
21. Embretson, S.E., 1998. A cognitive design system approach to generating valid tests. Application to abstract reasoning. Psychological Methods, 3: 380-396.
22. Embretson, S.E., (Ed)., 1985. Test design Developments in psychology and psychometrics. Orlando Academic Press.
23. Essentials of a Good Psychological Test. Retrieved on 10-06-2006 from http://www.wilderdom.com/personality/L3-2 Essentials Good Psychological Test.html.
24. Fleming, M., 2006. Evaluation and Assessment. Languages of Schooling towards a framework for Europe. Council of Europe, Language Policy Division, Strasbourg Retrieved from www.coe.int/lang.
25. Gotwal, A.W. and N.B. Songer, 2006. Cognitive Predictions Bio KIDS Implementation of the PADI Assessment System, (PADI Technical Report 10). Menlo Park, CA: SRI International.
26. Greaney, V. and T. Kellaghan, 2008. Assessing National Achievement Levels in Education. The World Bank. USA.
27. Greeno, J.G., A.M. Collins and L.B. Resnick, 1996. Cognition and learning. In D. C. Berliner and R. C. Calfee (Eds), Handbook of educational psychology. New York: Macmillan, pp: 15-46.
28. Gronlund, N.E. and R.L. Linn, 2005. Measurement and assessment in Teaching. New Delhi Baba Barkha Nath Printers.
29. Hambelton, R.K., 2000. Emergence of item response modeling in instrument development and data analysis. Medical Care, 38: 60-65.
30. Mathematics Framework for 2002. National Assessment of Educational Progress. National Assessment Governing Board. USA: Department of Education.
31. Mathematics Framework for 2003. National Assessment of Educational Progress. National Assessment Governing Board. USA: Department of Education.
32. McCallum, W.G., 2007. Assessing the Strands of Student Proficiency in Elementary Algebra. In: A.H. Schoenfeld, (Ed.). 2007. Assessing mathematical proficiency. Cambridge University Press. New York.