Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 2): pp. 1642-1658
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.02.184

Development Of Valid And Reliable Mathematics Achievement Test

Haleema Bano Ph.D Research Scholar, Department of Education, The University of Haripur

Muhammad Saeed Khan Associate Professor, Department of Education, The University of Haripur

Shaista Irshad Khan Assistant Professor, Department of Education, Abdul Wali Khan University Mardan

Abstract

This paper presents the efforts made to establish validity and reliability of a mathematics achievement test developed by researcher. The sample consisted of 80 students of Grade 7th, studying in two sections of an urban public sector school. Data of mathematics test scores were collected in two phases of pilot study. Two evaluation tools requiring SMEs (subject-matter experts) to rate test items on their alignment with ILOs of NCM 2006 and levels of Blooms Taxonomy, were also used to collect data for calculating content and construct validity. Item analysis was used for finding discrimination power and Cronbach Alpha used to calculate reliability value. The test is found moderate to highly reliable (0.79), content perfectly aligned (86.65 % of test items) with curriculum ILOs and well matched (K= 41.66 %, Com= 36.66 %, App= 21.68 %) with levels of Blooms Taxonomy and possesses good power of discrimination (Reasonably good=78 % test items, DI values 0.4-0.3, Marginally good=18 % test items, DI values 0.2-0.29).

Key Words: Test Development, validity and reliability, ILOs, discrimination power

Introduction

A valid and reliable measurement instrument underpins the credibility of entire research endeavor. Cronbach and Meehl (1955) identified the issue of validity for the first time while defining standard for evaluation of psychological tests. The validity pertaining to the content of test comes at first place. As asserted by AERA et al. (1999) validity is not an inherent feature of a test, rather is determined in relation to the purpose of that test. Robson (2011) takes test validity as degree to which a test can measure truly what it aims to gauge. Oliver (2010) attaches immense importance to the validity of research instrument irrespective of the type of research. Lodico, et al (2010) while exploring the concept of validity at length, placed content validity of the measuring tool, at the top using indicators of adequacy, relevance and appropriateness of content. Taherd oost (2016) also provided a detailed discussion about various forms of validity and reliability. More frequently identified forms of validity in literature are content, construct and concurrent validity. Creswell (2005) defines content validity of a test as the ability of test items and their resulting scores to represent all possible items and their scores in that content domain.

Half a century ago, concern regarding test item scrutiny emerged to monitor if the items are free from any gender, racial or academic biases and later grew to assure quality in terms of validity evidence (Gomez-Benito, et al., 2018). Sireci and Falkner-Bond (2014) perceived test validity in a broader perspective of test justification for attainment of certain ends. Lane (2014) connects test validity with the purpose of testing as better educational outcomes. The gap between observed and expected differences in scores on a test, due to group diversity is measured as differential item function (DIF), and is statistically calculated as content validity evidence (Gomez-Benito, et al., 2018). Ensuring validity of instrument is a basic concern while using research evidence from experimental studies (Kane, 2006; Linn, et al., 2010). It is important to understand strength and weaknesses of an assessment plan before using test scores effectively (Krell & Hui, 2017). While discussing the concept of content validity at length, Sireci et al., (2008) introduced the terms of domain definition, domain representation, domain relevance and appropriateness of test construction procedures. What helps define domain is the clear and objective explanation of the content areas and abilities that test aims to measure through it, the test blue print and intended learning outcomes for that content as mentioned by curriculum. Sireci and Faulkner-Bond (2014) explained how sufficiently a test represents the content that it aims to cover is its domain representation attribute. It is the job of subject matter experts to examine and weigh the test to see if the content has been thoroughly and fully addressed in test development (Crocker, Miller & Franks, 1989) which more recently is termed as test alignment research (Bhola, et al., 2003) where congruency of test items with curriculum framework is gauged.

Very few writings in education, have enjoyed the status of being most extensively read, used and referred document by educators across the globe over more than half a century, as is the case of Bloom's Taxonomy (Anderson & Sosniak, 1994; Anderson & Krathwohl, 2002; Marzano & Kendall, 2007). Armstrong (2016) acknowledges the contribution of Blooms taxonomy as a most frequently used tool by teachers for assessment purposes. Since then, it has been used for more structured and objective-driven assessment in education (Marzano & Kendall, 2007). Adam's (2015) points out two important functions that Bloom's Taxonomy can serve for educators. It enables teachers to design objectives in outcome or performance form i.e., observable thus making objective assessment, possible. Secondly it sensitizes educators to gradually move towards engaging higher order thinking of learners for deep understanding. Forehand (2010) perceived Bloom's Taxonomy as a graded pathway to proceed towards deeper thinking. In the present study the Blooms taxonomy, in its original form is used as the underpinning construct of mathematics curriculum in designing the intended leaning outcomes.

Reynolds and Kearns (2017) define backward planning model as to formulate first, the desired outcomes then figure out the assessment appropriate for gauging that outcome and lastly to plan such instructional strategies that may help achieve those outcomes. The entire course of planning keeps the learner at pivotal position. The National Curriculum for Mathematics (2006), content source for the research tool, presents the content through backward course design. This paper presents the alignment of test items with the relevant outcomes, mentioned in the NCM (2006). Kelting-Gibson (2003**)** refers to the Covey's (1998) rational for backward course planning "To begin with the end in mind means to start with a clear understanding of your destination. It means to know where you're going so that you better understand where you are now so that the steps you take are always in the right direction". Wiggins and Mc Tighe (1998) presented a changed sequence of steps taken for curriculum planning; "1) identify the desired results, 2) determine the acceptable evidence, and 3) plan learning experiences and instruction".

Wiggins and Mc Tighe (2005) while summarizing the benefits of backward course planning, state that teacher can present most relevant content, effectively utilize time, better plan presentation, get students more involved in learning and provide feedback more often on learners' progress. Backward lesson plan is an outcome driven strategy which focuses on what learners will ultimately gain in terms of knowledge and skills (Chizhik & Chizhik 2018). While criticizing learning objectives, Mc Tighe and Seif (2003) found backward course design successful in improving learner's educational outcomes. At higher secondary level such model facilitates higher order thinking (Trigwell, 2010; Wang et al., 2014) and more satisfactory learning in the science subjects is witnessed as compared to the traditional content-driven course planning (Wood, 2009; Bauer-Danto in, 2009; Dolan & Collins, 2015). Recent research offers a good deal of knowledge about structure, techniques and benefits of backward course planning (Trigwell, 2010; Singer et al., 2012; Baker et al., 2014).

Developing Mathematics Achievement Test

There are eight developmental stages summarized by Icebacak and Ersoy (2017) from literature on developing achievement test.

- The Area to be used for Test Scores,
- Determining the behaviors representing the area or the statement,
- Writing Test İtems,
- Reviewing The Test İtems,
- Preparing The Test Form,
- Putting The Test on A Trial Implementation,
- Selecting Materials by Analyzing Them According to The Trial Implementation,

- Prognosis of the selected items that generate the statistics of the final test

Assessment is a vital element of entire course of study (Bayrak & Erden, 2007). Assessment brings into light, the most important information of learners' attainment of objectives and the degree of gaining knowledge, skills and attitude (Sönmez & Alacapınar, 2013). Assessment reveals the effectiveness of learning program. It has implications for the significance of assessment tool. The multiple-choice test is referred as the most frequently used tool for assessment purpose (Kempa, 1986;Bekiroğlu, 2004) as well as a tool to create evidence for effectiveness of experimental studies in education (Incebacek & Erosy 2017). The assessment tool, however must fulfill the criterion of validity and reliability (Çelikler & Kara, 2015; Belgin & Esen, 2017; Reena & Anisha, 2017).

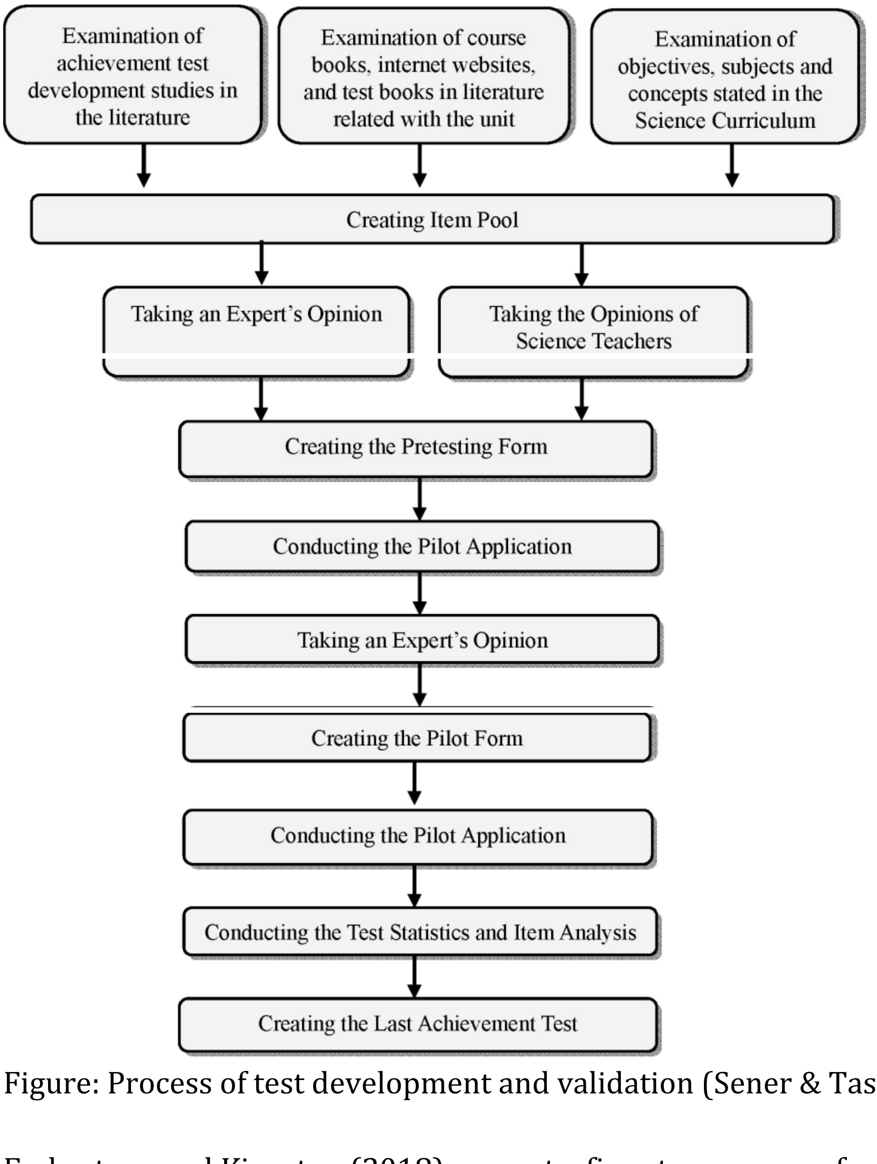
Examination of achievement test development studies in the literature
Examination of course books, internet websites, and test books in literature related with the unit
Examination of objectives, subjects and concepts stated in the Science Curriculum
Creating Item Pool
Taking an Expert's Opinion
Taking the Opinions of Science Teachers
Creating the Pretesting Form
Conducting the Pilot Application
Taking an Expert's Opinion
Creating the Pilot Form
Conducting the Pilot Application
Conducting the Test Statistics and Item Analysis
Creating the Last Achievement Test

Figure: Process of test development and validation (Sener & Tas, 2017)

Embertson and Kingston (2018) suggest a five stage process for developing a valid and reliable achievement test. Once the test blueprint is prepared, the first step is to select most appropriate item developers and then train them. Writing test item is second stage. Third stage requires monitoring alignment of items with test blue print carried out by subject matter experts. After incorporating experts' feedback, the test items are now pilot tested to seek imperial evidence of validity, reliability, distinctive power and difficulty level. The fifth stage involves the compilation of finalized items into a test based on the results of pilot testing.

Mardapi (2012) suggests nine stages for test construction where as Pandra, Sugima and Mardapi (2017)made it ten stage process. The most significance stage, according to Brennan (2006) however is the decision making about the purpose, context and content of the test.

The strategic layout of test development procedures remains the same, except keeping it brief in few stages i.e., five stages (Embertson & Kingston, 2018), or giving a detailed process of more stages i.e., eight stages (Icebacak & Ersoy, 2017) or ten stages (Pandra, Sugima and Mardapi, 2017). Researcher has followed the test development process suggested by Sener and Tas (2017).

Research Question

This paper aims to investigate following research questions:

i. Whether the researcher developed Mathematics Achievement Test is aligned with the Curricular ILOs mentioned for Grade six and seven?
ii. To what extent the distribution of test items in various levels of Blooms Taxonomy matches the construct?
iii. How reliable is the test for evaluating mathematics achievement of seventh graders?
iv. How powerful are the test items in discriminating high achievers from low achievers?

RESEARCH METHODOLOGY

Sample

A total of eighty female students studying in grade seven in urban public high school participated in the pilot testing of mathematics achievement test, forty in first pretest piloting and forty in the second pilot study.

Data Collection Tool

Two evaluation tools were developed by researcher, for subject matter experts to provide evidence for content and construct validity. In Evaluation Sheet I, test items were presented alongside six levels of blooms Taxonomy and experts were asked to mention the level relevant to each item (construct validity). In Evaluation Sheet II, experts provided feedback on whether test items are aligned with curriculum ILO's (content validity) wherein items were presented against the curricular ILO's (Intended Learning Outcomes) and experts were asked to rate them on a five-point scale of Not

Aligned (0) to Perfectly Aligned (4). A multiple choice mathematics achievement test was used to collect data during two phases of pilot study.

Multiple Choice Questions based Mathematics Achievement Test (MAT)

Lodico, et al (2010) preferred using a test developed by researcher, arguing that a standardized test available with the title appearing similar to the one of our study, may not necessarily be suitable for the population we aim to study. For instance, the readability level of our population may be different/ low than expected. Similarly, assumption about prior knowledge of population may not be true for population of our study. Hence researcher decided to develop the tool by herself. This research tool consisted initially, of eighty multiple choice questions developed from chapter number seven to thirteen (second half) of Mathematics Textbook for grade VI and VII (Khyber Pakhtunkhwa Text Book Board, Peshawar), covering five content strands including:

- Financial Arithmetic
- Algebraic Expression
- Linear Equations
- Fundamentals of Geometry
- Practical Geometry
- Circumference, Area and Volume
- Information Handling.

Chapter-wise percent weightage evidence from National Curriculum for Mathematics (2006)

Unit	Title Grade VI	NCM %age GradeVI	Test Items %age Grade VI	Title Grade VII	NCM % age Grade VII	Test Items % age Grade VII
7	Financial Arithmetic	5	**5**	Financial Arithmetic	**7**	**10**
8	Introduction to Algebra	7	**5**	Algebraic Expression	**10**	**10**
9	Linear Equation	8	**5**	Linear Equation	**5**	**7**
10	Geometry	15	**10**	Fundamentals of Geometry	**12**	**15**
11	Perimeter and Area	7	**8**	Practical Geometry	**15**	**15**
12	Three Dimensional Solids	8	**5**	Circumference Area and Volume	**8**	**10**

13	Information Handling	5	**2**	Information Handling	**5**	**3**
Total		55 %	40 %		**60 %**	**70 %**

Mathematics achievement test purposively included test items from both Grades, VI (21 items= 35 % of the test content) and VII (39 items= 65 % of test content) due to the fact that same tool was to be used for pre and posttest evaluation. Unit number 7-13 of Grade VI and VII (with same unit titles), as per Curricular percentage allocation (NCM, 2006, p. 141), requires 55 % and 60 % respectively, of the content be taken for tool whereas this test included 40 % and 70 % of the content as test items from the second halves of the books respectively. Difference in the prescribed versus used percentage is due to the use of same tool for pre and post intervention evaluation.

Keeping in view the Assessment standards of National Curriculum for Mathematics (2006, p.137), the test items required students to comprehend, analyze, evaluate, discriminate and reason mathematically.

Results and Analysis

The tool was first piloted on 40 students other than expected participants of study, on February 22nd 2020. Objective scoring using answer key was done, data tabulated in excel and analyzed using Kolmogorov-Smirnov and Shapiro-Wilks test for **normality,** and found normally distributed.

Discrimination power of an item, an indicator of its quality (Ebel & Frisbie, 2004), was calculated for all eighty items. The results produced 30 items with no discrimination power (DI Value: below 0, eliminated from test), 29 items as weakly discriminating high achievers from low achievers (DI Value: 0-0.18, revisions made in content and phrasing),11 items as moderately good discriminator (DI Value: 0.27), 8 items as reasonably good discriminators (DI Value: 0.36) and 2 items as very good discriminator of high and low achievers.

Discrimination Index Values of Test Items Based on 1st Pilot Study Results

Item Deleted DI Value 0 to -0.18		Poor Items DI Value 0.09 to 0.18	Marginally Good Items DI Value 0.27	Reasonably Good Items DI Value 0.36	Very Good Items DI Value 0.4 and above	Total Items

	80 (-0.09)
	79 (-0.04)
	76 (-0.09)
		7 (0.18)
		11 (0.09)
		12 (0.09)
		13 (0.18)
		17 (0.09)
		19 (0.09)
			8 (0.27)
			10 (0.27)
			15(0.27)
			23(0.27)
			37(0.27)
			40(0.27)
				73 (0.36)
				64(0.36)
				60 (0.36)
				50(0.36)
				46(0.36)
				31(0.36)
					16 (0.45)
					61 (0.54)	

	74 (-0.18)
	70 (-0.09)
	67 (0)
	59 (-0.09)
	57 (0)
	56 (-0.18)
	49 (-0.09)
	48 (0)
	47 (-0.18)
	45 (0)
	43 (-0.09)
	42 (0)
	38 (0)
	36 (0)
	33 (0)
	30 (0)
	28 (-0.09)
	27 (-0.09)
	26 (0)
	25 (-0.18)
	22 (0)
	18 (-0.18)
	14 (0)
	9 (0)
	6 (-0.09)
	5 (0)
	4 (0)
		20 (0.09)
		21 (0.18)
		24 (0.09)
		29 (0.09)
		32 (0.09)
		34 (0.18)
		35 (0.18)
		39 (0.18)
		41 (0.18)
		44 (0.18)
		51 (0.18)
		52 (0.18)
		53 (0.18)
		58 (0.18)
		62 (0.09)
		63 (0.09)
		65 (0.09)
		66 (0.09)
		68 (0.09)
		69 (0.18)
		71 (0.09)
		72 (0.09)
		75 (0.09)
			54(0.27)
			55(0.27)
			77(0.27)
			78(0.27)
			3 (0.27)
				2 (0.36)
				1 (0.36)		

Total	30	29	11	8	2	80

After major revisions made (thirty items eliminated, twenty-nine revised both phrasing and content, eleven items rephrased and twenty new items developed based on experts' feedback), the new test comprised of 60 items. This tool was piloted once again with another section of class seventh (45 students) in public sector girls high school other than sample institution, on March 25th 2020.

The scores of 2ndpilot testing were analyzed using Kolmogorov-Smirnov and Shapiro-Wilks test for normality, and found normally distributed. Discrimination Index Values found for all sixty items of the revised test. Twelve out of sixty items (20%) were very good (DI equal and above 0.4), thirty-five items (58.33%) were reasonably good (DI between 0.3 and 0.39) eleven items(18.33%) were marginally good (DI between 0.2 and 0.29) whereas only two items were found as poor (DI between 0.1 and 0.19). The discrimination index used to evaluate test items was proposed by Ebel and Frisbie (2004, p. 232).

Discrimination Index Values of Test Items Based on 2nd Pilot Study Results

Number of Items (% age in test)	Discrimination Index Value	Status of the Item
12 (20)	≤0.4	Very good
35 (58.33)	0.3≤0.39	Reasonably good
11 (18.33)	0.2≤ 0.29	Marginally good
2 (3.33)	0.1≤ 0.19	Poor

Validating Mathematics Achievement Test

Two evaluation tools were developed by researcher, for subject experts to provide evidence for content and construct validity. In Evaluation Sheet I, test items were presented alongside six levels of blooms Taxonomy and experts were asked to mention the level relevant to each item (construct validity). The percent calculated for each domain revealed that 41.66 % items fall in knowledge domain, 36.66 % in comprehension domain and 21.68 % in application domain.

Evidence of Construct Validity: Test Items Alignment with Levels of Blooms Taxonomy

NCM Standard	% of Test (No. of Items)	% Items in Knowledge Domain (No. of Items)	% Items in Comprehension Domain (No. of Items)	% Items in Application Domain (No. of Items)
Standard 1 Numbers and Operations	15 % (09)	5% (03)	3.33% (02)	6.66% (04)
Standard 2 Algebra	28.33 % (17)	15% (09)	8.33% (05)	5% (03)
Standard 3 Measurements	(51.66 %) 31	18.33% (11)	23.33% (14)	10% (06)

and Geometry				
Standard 4 Information Handling	5 % (3)	3.33% (02)	1.66% (01)	-
Total	60	41.66 %	36.66 %	21.68%

In Evaluation Sheet II, experts provided feedback on whether test items are aligned with curriculum ILO's (content validity) wherein items were presented against the curricular ILO's (Intended Learning Outcomes) and experts were asked to rate them on a five-point scale of Not Aligned (0) to Perfectly Aligned (4). Vakili and Jahangiri (2018) mentions minimum of five experts for feedback on validity of research tools. Researcher approached eleven experts out of which ten experts provided feedback.

Experts included Additional Director, Directorate of Professional Development KP, three Instructors Regional Institute for Teacher Education Mardan (Male), Head Department of Mathematics Kakul Academy, Lecturer Mathematics Kakul Academy, Principal Government Girls High School Mardan, Principle Government Higher Secondary School Mardan and two Government School Teachers. The data analysis of experts' feedback revealed four test items (item number: 22,26,31,48) (3.33%) out of sixty were poorly aligned with Intended Learning Outcomes mentioned in National Curriculum for mathematics (2006), thus eliminated from test and new items included, four items (2,7,13,60) (3.33%) were found somewhat aligned, were revised and fifty-two items (86.65%) were found perfectly aligned with ILO's of NCM (2006).

Evidence of Content Validity: Test Items Alignment with ILO's (NCM, 2006)

NCM Standard	% of Test (No. of Items)	% of items rated as Poorly Aligned	% of items rated as Somewhat Aligned	% of items rated as Perfectly Aligned
Standard 1 Numbers and Operations	15 % (09)	-	3.33% (2 Items: 2,7)	11.66 % (7 Items)
Standard 2 Algebra	28.33 % (17)	3.33 % (2 Items: 22, 26)	1.66 % (1 Item: 13)	23.33 % (14 Items)
Standard 3 Measurements and Geometry	51.66 % (31)	3.33 % (2 Items: 31, 48)		48.33 % (29 Items)
Standard 4 Information Handling	5 % (3)		1.66 % (1 Item: 60)	3.33 % (2 Items)
Total	60	6.66%	6.65 %	86.65 %

Test blue print when compared with experts' feedback, provided the evidence for relevance, adequacy and appropriateness of the test, establishing its content validity and construct validity.

The concurrent validity (comparability of test results with another standardized tool to see if they are correlated and is the test capable of replacing that standardized tool and predictive validity (test scores found correlated with another achievement test taken at later point of time) were not considered for this achievement test.

Valid instrument has to be reliable as well but a reliable instrument may not be valid necessarily (Thatcher, 2010; Twycross & Shields, 2004).

Reliability of Research Instrument

Joppe (2000) defines reliability as accuracy, replicability and dependability of an instrument. Creswell (2005) refers it as consistency and stability of the scores produced by a tool. Researcher must know the degree to which the tool is reliable (Huck, 2007). Reliability of research instrument can be seen in two different dimensions: 1) being internally consistent and coherent and 2) be able to yield same results on repeated measures. The former is concerned with ability of various test items to measure the one same concept and later reflects the stable and sound quality of test items (Muijs, 2004; Mohajan, 2017). Tool used in this study was evaluated for internal consistency of the items as measure of reliability.

Method used to Establish Reliability

Internal consistency of a test can be gauged either by using split-half method or coefficient alpha (Muijs,2004). There are more than one ways to measure reliability (Muijs, 2004) of which Cronbach Alpha is the most commonly used (Taherdoost, 2016). For this study split-half method was used. Test scores for all even numbers of items were compared with the test scores of all odd number of test items to see how close their relation is. The value obtained for reliability coefficient was 0.79. Downing (2004) suggest reliability coefficient as high as above 0.9 for tools used in financial and professional decision making, however he considers above 0.7 as acceptable for research tools. For Brown (2002) this tool is 79 % reliable. Hence this mathematics achievement test may be referred as sufficiently/moderately reliable.

Conclusion

This mathematics achievement test (comprised of 60 items) is found reasonably good discriminator of high achievers from low achievers (78 % items) with average DI value of 0.4 according to index used by Ebel and Frisbie (2004). The reliability coefficient refers this test as moderate/sufficiently reliable instrument. Bal-Incebacak and Erosy (2017) found their test (comprised of 43 items) a strong discriminator with average discrimation value of KR-21 0.8. Downing (2004) considers this reliability as acceptable for research tools. A research tool with KR-20 coefficient value between 0.7-0.8 is considered (Buyukozturk, 2008; Fraenkel, Wallen & Hyun, 2011) as sufficiently reliable.

Bal-Incebacak and Erosy (2017) found the reliability coefficient as high as 0.9, of a mathematics achievement test for fourth graders on topic of Fraction. A similar study conducted by Sener and Tas (2017) found their Biology test for Grade Five (comprised of 38 items) having KR-20 reliability coefficient 0.87, very good discrimination power with DI value of 0.49. Reliability of a third grade mathematics achievement test (comprised of 29 items) developed by Pandra et al., (2017) was found good with reliability coefficient value of 0.78 and average discrimination power was found satisfactory.

A high percentage of test items is found perfectly aligned with National Curriculum for Mathematics suggested ILO's. The alignment of test items with various levels of Bloom Taxonomy revealed that majority of test items fall in knowledge domain, second highest in comprehension domain and next lower percentage falls in application domain. It may be concluded that the ratio between knowledge, comprehension and application domain is approximately 10: 9: 5.

References

Adams, E. N. (2015). Bloom's taxonomy of cognitive learning objectives. Journal of the Medical Library Association, 103(3). 152-153. Retrieved from https://www.ncbi.nlm.nih.gov

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2002). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York, NY: Longman.

Anderson, L. W., & Sosniak, L. A. (Eds.). (1994). Bloom's taxonomy: A forty-year retrospective. ninety-third yearbook of the national society for the study of education. Chicago, Chi: University of Chicago Press.

Armstrong, P. (2016). Bloom's taxonomy. Retrieved from https://evawintl.org/wp-content/uploads/Blooms-Taxonomy.pdf

Baker, L. A., Chakraverty, D., Columbus, L., Feig, A. L., Jenks, W, S., Pilarz, M., Wesemann, J. L. (2014). Cottrell scholars collaborative new faculty workshop: Professional development for new chemistry faculty and initial assessment of its efficacy. Journal of Chemical Education, 91(11), 1874–1881. Retrieved from https://pubs.acs.org/doi/abs/10.1021/ed500547n

Bauer-Dantoin, A. (2009). The Evolution of Scientific Teaching within the Biological Sciences. In R. A. R. Gurung, N. L. Chick, & A. Haynie (Eds.), Exploring Signature Pedagogies: Approaches to Teaching Disciplinary Habits of Mind, (pp. 224–43). Sterling, VA: Stylus.

Bayrak, B., & Erden, A. M. (2007). The evaluation of science curriculum. Kastamonu Education Journal 15(1), 137-154.

Bekiroğlu, O. (2004). How successful classical and alternative measurement - evaluation methods and applications in physics. Ankara: Nobel Publication Distribution.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. National Council on Measurement in Education, 22(3), 21-29. Retrieved from https://doi.org/10.1111/j.1745-3992.2003.tb00134.x

Brennan, R. L. (2006). Educational measurement. Iowa City: United State of America: American Council on Education and Praeger Publisher

Brown, J. D. (2002). The Cronbach alpha reliability estimate. Shiken: JALT Testing & Evaluation, 6(1), 17-18. Retrieved from https://hosted.jalt.org/test/PDF/Brown13.pdf

Celikler, D., & Kara, F. (2015). Development of achievement test: Validity and reliability study for achievement test on matter changing. Journal of Education and Practices, 6(24). Retrieved from https://eric.ed.gov/?id=EJ1078816

Chizhik, W, E., & Chizhik, W. A. (2018). Using activity theory to examine how teachers' lesson plans meet students' learning needs. The Teacher Educator, 53(1), 67-85 Retrieved from https://www.tandfonline.com/doi/abs/10.1080/08878730.2017.1296913

Covey, S. R. (1989). The 7 habits of highly effective people. New York, NY: Simon & Schuster.

Creswell, J. W. (2005). Educational research: Planning, conducting, and evaluating quantitative and qualitative research. Upper Saddle River, New Jersey, NJ: Pearson Education.

Crocker, L. M., Miller, D., & Franks E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. Applied Measurement in Education, 2(2), 179-194. Retrieved from https://www.tandfonline.com/doi/abs/10.1207/s15324818ame0202_6

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302. Retrieved from https://psycnet.apa.org/record/1956-03730-001

Dolan, E. L. & Collins, J. P. (2015). We must teach more effectively: Here are four ways to get started. Molecular Biology of the Cell, 25, 2151–2155. Retrieved from https://www.molbiolcell.org/doi/abs/10.1091/mbc.E13-11-0675

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. Medical Education, 38(9), 1006-1012. Retrieved from https://doi.org/10.1111/j.1365-2929.2004.01932.x

Ebel, R. L. & Frisbie, D.A. (2004). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Embretson, E. S., & Kingston, M, N. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? Journal of Educational Measurement, 55(1), 112-131. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jedm.12166

Forehand, M. (2010). Bloom's Taxonomy: From emerging perspectives on learning, teaching and technology. The University of Georgia. Retrieved from https://www.d41.org/cms/lib/IL01904672/Centricity/Domain/422/BloomsTaxonomy.pdf

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2011). Single-subject research. In J. R. Fraenkel, N. E. Wallen, & H. Hyun (Eds.), How to design and evaluate research in education (pp. 301–329). New York, NY: McGraw-Hill Companies

Gomez-Benito, G. J., Sireci, S., Padilla, L. J., Hidalgo, D. M., & Benitez, I. (2018). Differential item functioning: beyond validity evidence based on internal structure. Psicothema, 30(1). 103-117

Government of Pakistan. (2006). National curriculum for mathematics. Ministry of Education. Islamabad: Ministry of Education.

Huck, S. W. (2007). Reading statistics and research. United States of America, USA: Allyn & Bacon.

Incebacek, B. B., & Ersoy, E. (2017). Developing an achievement test for fraction teaching: Validity and reliability analysis. ITM Web Conference, 13(1), 1-14 Retrieved from https://www.itm-conferences.org/articles/itmconf/abs/2017/05/itmconf_cmes2017_01003/itmconf_cmes2017_01003.html

Joppe, M. (2000). The research process. Retrieved from https://www.uoguelph.ca/hftm/1-problem-definition

Kane, M. T. (2006). Validation. In B.L. Robert (Ed.), Educational Measurement (pp. 17-64). Westport, CT: Praeger.

Kelting-Gibson, L. M. (2003). Preservice teachers' planning and preparation practices: A comparison of lesson and unit plans developed using the backward design model and a traditional model[Unpublished doctoral thesis), Montana State University.

Kempa, R. (1986). Assessment in science. London, Cambridge: London Cambridge University Press.

Krell, M., & Hui, F. K. S. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. Cogent Education, 4(1). 1-10 Retrieved from https://www.tandfonline.com/doi/full/10.1080/2331186X.2017.1280256

Lane, S. (2014). Validity evidence based on testing consequences. Psicothema, 26(1), 127-135. Retrieved from https//doi.org/10.7334/psicothema2013.258

Laudico, M. G., Spaulding, D. T., &Voegtle, K. H. (2010). Methods in educational research: From theory to practice (2nd ed.). San Francisco, CA: Jossey-Bass.

Linn, M. C., Chang, H. Y., Chiu, J. L., Zhang, Z. H., & Mc Elhaney, K. (2010). Can desirable difficulties overcome deceptive clarity in scientific visualizations? In A. Benjamin (Ed.), Successful ¬remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork (pp. 239–262). New York, NY: Routledge.

Marzano, R., & Kendall, J. (2007). The new taxonomy of educational objectives (2nded.). Thousand Oaks, CA: Corwin Press.

McTighe, J. & E. Seif. (2003). A summary of underlying theory and research base for understanding by design. Association for Supervision and Curriculum Development. Retrieved from http:// assets.pearsonschool.com/asset_mgr/current/201032/ubd_ myworld_research.pdf

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability.

Muijs, D. (2004). Doing quantitative research in education with SPSS. Sage Publications. Retrieved from https://dx.doi.org/10.4135/9781446287989

Pandra, V., Sugiman., & Mardapi, D. (2017). Development of mathematics achievement test for third grade students at elementary school in Indonesia. International Electronic Journal of Mathematics Education, 12(3), 769-776. Retrieved from https://www.iejme.com/article/development-of-mathematics-achievement-test-for-third-grade-students-at-elementary-school-in

Reena, R. & Anisha. (2017). Construction and standardization of mathematics achievement test for 9th grade students. Educational Quest, 8(3), 629-633. Retrieved from https://www.indianjournals.com/ijor.aspx?target=ijor:eq&volume=8&issue=3&article=027

Robson, S. (2011). Producing and using video data with young children: A case study of ethical questions and practical consequences. In Harcourt, D., Perry, B., & Waller, T. (Eds.), Researching young children's perspectives(pp. 178–192). London: Routledge.

Sener, N., & Tas, E. (2017). Developing achievement test: A research for assessment of 5th grade biology subject. Journal of Education and Learning, 6(2). 254-271. Retrieved from https://eric.ed.gov/?id=EJ1139232

Singer, S. R., Nielsen, N. R., & Schweingruber, H. A. (Eds.). (2012). Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. Washington, WA: National Academies Press.

Sireci, G. S., Baldwin, P., Martone, A., Zenisky, L. A., Kaira, L., Lam, W. Shea, C.L., Han, K.T., Deng, N., Delton, J., & Hambleton, K, R. (2008). Massachusetts adult proficiency tests technical manual. Center for Educational Assessment, Retrieved from https://www.umass.edu/remp/docs/MAPT_TechManual_v2.pdf

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema, 26(1). 100-107.

Sönmez, V. & Alacapınar, F. G. (2013). Illustrated scientific research methods (2ndeded.). Ankara: Anı Publisher.

Taherdoost, H. (2016). Validity and reliability of the research instrument; How to test the validation of a questionnaire/survey in research. SSRN Journal.5(3), 28-36 Retrieved from https://www.google.com/search?q=ssrn+journal&sxsrf=AOaemvIQGw-

Thatcher, R. W. (2010). Validity and reliability of quantitative electroencephalography. Journal of Neurotherapy, 14(2), 122-152. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/10874201003773500

Trigwell, K. (2010). Promoting effective student learning in higher education. In P. Peterson, E. Baker, & B. Mc Gaw, (Eds.) International Encyclopedia of Education,4, (pp 461–466) Oxford: Elsevier.

Twycross, A., & Shields, L. (2004). Validity and reliability--what's it all about? Part 1 validity in quantitative studies. Paediatric Nursing, 16(9), 28-29. Retrieved from https://www.proquest.com/openview/143a3a051b92c6843ea5ed955918b2b8/1?pq-origsite=gscholar&cbl=33983

Vakili, M. M., & Jahangiri, N. (2018). Content validity and reliability of the measurement tools in educational, behavioral, and health sciences research. Journal of Medical Education Development, 10(28). 106-118. Retrieved from https://zums.ac.ir//edujournal/article-1-961-en.html

Wang, X., Su, Y., Cheung, S., Wong, E., & Kwong, T. (2014). An exploration of Biggs' constructive alignment in course design and its impact on students' learning approaches. Assessment and Evaluation in Higher Education,38(4), 477–91.

Wiggins, G., & McTighe, J. (1998). Understanding by design. Alexandria, VA: Association for Supervision and Curriculum Development.

Wood, W. B. (2009). Innovations in teaching undergraduate biology and why we need them. Annual Review of Cell and Developmental Biology, 25(1), 93–112. Retrieved from https://www.annualreviews.org/doi/abs/10.1146/annurev.cellbio.24.110707.175306robson