JIES

Journal of Interdisciplinary Educational Studies
October 2023, Vol. (03) Issue 2, 47-57

Developing and Validating Item Bank in Science and Mathematics at Primary Level Using Item Response Theory

Misbah Shahid
Former M. Phil Scholar,
Institute of Education & Research, University of the Punjab, Lahore.
bmisbahshahid@gmail.com

Muhammad Saeed
Professor & Associate Dean,
Faculty of Social Science and Humanities, Minhaj University Lahore.
drsaeed1961@hotmail.com

Mubashara Akhtar
Assistant Professor,
Faculty of Education, LCWU, Lahore.
mubashara.akhtar@lcwu.edu.pk

Abstract

An item bank is relatively a large collection of calibrated items categorized by subject, item type and grade level. This quantitative study provides the procedure for developing, assembling, validating and calibrating item bank using item response theory. The researchers developed 300 items, 150 in the subject of Science and 150 in the subject of Mathematics that was split into two tests, each of 75 items. Each test was administered on 960 students of Grade 5. The content of the tests was based on the national curriculum (2006) and textbooks of Science and Mathematics of Grade 5. The tests were piloted on 50 students in two urban public schools of district Lahore. Accordingly, test items were reviewed and improved. Table of specification was developed for both subjects Science and Mathematics. Both tests were validated by consulting three relevant experts. After collecting the data, items were analyzed by using ConQuest software. Total 300 items were analyzed 150 in the subject of Science and 150 in the subject of Mathematics. Total 224 test items were fulfilling the criteria of item difficulty (0.11 to 0.91) and discrimination index (0.20 to 0.40), out of which 113 in the subject of Science and 111 in the subject of Mathematics. The study recommended to use the developed items of Science and Mathematics on a large scale to measure the abilities of the students. Teachers may prepare their students of Grade 5 for Punjab Examination Commission (PEC) by using these items.

Key words: Item bank, Item response theory, Item calibration

Introduction

At all levels of education today, the need of quality assessment is one of the essential component for successful education system. The central question is how do we determine quality assessment? Different kinds of technology and methods are being adopted by teachers and

institution to promote quality assessment but "item bank" is the most popular one. Item bank is a way to measure predetermined set of constructs that is based on the development of catalog of items (Stoeger, 2017). An item that is mostly known as questions is used to solve problems and adequately provide the solution to specific questions. Classifying items in to different categories is the basic craft of item banking. Creation of valid and reliable content and different forms of test is the main function of item bank that ultimately supports high quality assessment. Numerous purposes are performed by high quality assessment for students it provides the foundation for continuing their study to a higher level of education and moreover helps them to look for better jobs. Success of institution and individual student are based on the evidence of assessment results (Friyatmi et al., 2020).

Modern and quality assessment is depending on item banks. Item bank is a kind of database in which different types of questions are stored after measuring the difficulty and discrimination index and then it was coded by item type, grade level and subject area (Gronlund, 1998). Item banks have made testing easier, effective and faster and it likely to be very beneficial for test developers and teachers. In 1960s, the movement of individualized instruction and behavioural objectives are linked with the idea of item bank. Item bank is a new idea in test construction and test advancement (Van der Linden, 1986). Item bank has reduced the burden of teachers in test planning and test advancement because item banks contain large collection of good items. For instance, the quality of test used in schools, could be better than those tests without item banking. Testing program will be more adequate and flexible when item banks are developed on the concept of item response theory. As item response theory helps to compare the results on same standard even when different groups of students can take different tests.

Psychometric properties of items are adequately provided by item banks. Psychometric was explained by two approaches such as traditional and modern approach. Classical testing theory was the conventional access. It is centered upon different forms of validity, test-retest reliability, standardization, internal consistency and normative data. Item response theory is the latest approach or technique used for item calibration. It focuses on how adequately test item functions in accessing constructs. The most important function of item response theory is to develop parallel form of tests, offer adaptive computerized testing and to scale test items for difficulty and item discrimination (DeMars, 2010). Individual items of a test was the fundamental concept of item response theory rather than the accumulation of item response such as test score (Baker, 2001). As new technologies are incorporated in the field of education, testing program also start to switch from conventional approach to modern approach such item response theory. Incorporation of technology in designing a test, the test development more sophisticated. Great flexibility in scoring and scaling the accessed traits becomes easy with the use of item response theory as compared to classical testing theory. In the implementation of computerized adaptive test the item independent scoring in item response theory is necessary. Item calibration was the most important part of test operation when item response theory was used to model test data. The method of collecting responses from a sample of examinee by fitting item response theory models and then estimating the items parameters using the required data is known as item calibration.

Large item bank was the new approach to promote modern assessment because the items we take from item banks to develop test was calibrated. The reliability and validity of the test is directly related to the accuracy of the calibrated item parameters such as item difficulty and discrimination index. Software that is based on the IRT methodology is still user-unfriendly as related to many commercially offered statistical packages. Analyzing data by using IRT methodology is much more challenging as compared to classical testing theory (Kline, 2005).

Before adding items in item bank, it has to be calibrated. The method of collecting responses from a sample of examinee by fitting IRT model and then estimating the item parameter using the data is known as item calibration. Validity and reliability of the test was directly effects the accurateness of calibrated item parameters. Scholastic achievement of the students in most of the schools was measured by standardized tests. To avoid administering the similar test at each year these schools have to revise tests with some frequency. Generating new tests was always a very time taking process. Test developer job was not confined to the composing of test items but one of the most important job of test developer was to determine the difficulty level of each item to guarantee that a test was neither too easy nor too tough. Test developers can escape this lengthy method by using item banks. (Lawrence, 1998).

In Pakistan such kind of examples are rarely found. National Testing Service (NTS) is one of the organization that conduct tests for student's admissions and scholarships, and employs' recruitment and promotion purposes of various organization. There are 77 disciplines in which NTS stored more than 600,000 items in the subject of Medicine, Management, Natural or Social Sciences, IT and Engineering. But these items are not truly developed following IRT. Furthermore, the validity and reliability of items is questionable as they are not piloted on an adequate sample. This study will be an attempt to develop and validate an item bank of Science and Mathematics at grade 5 using item response theory.

Objectives of the Study

Following were the research objectives of the study:

1. Develop items in the subject of science at Grade 5 level using item response theory.
2. Validate items of science at Grade 5 level using item response theory.
3. Develop items in the subject of mathematics at Grade 5 level using item response theory.
4. Validate items of mathematics at Grade 5 using item response theory.

Research Questions

Based on the above objectives following research questions were posed.

1. What will be the viable steps for the development of item bank?
2. How far developed items using IRT can assess lower cognitive abilities of students in the subject of science at grade 5 level?
3. How far developed item bank of science at grade 5 is valid and reliable?
4. How far developed items using IRT can assess lower cognitive abilities of students in the subject of mathematics at grade 5 level?
5. How far developed item bank of mathematics at grade 5 is valid and reliable?

Methodology

The study was quantitative in nature. Quantitative in terms of that a test of science and mathematics was solved by the students of grade 5. The researchers developed 150 items of science and 150 items of mathematics for grade 5 using item response theory and textbook of science and mathematics. And for the validation of the items, a test was developed that was

validated by one university teacher and two subject teachers. A total of 300 Multiple choice questions was calibrated using item response theory. The psychometric properties of each item were optimizing. Researchers used multistage sampling technique for the collection of data from students. After analysis, researchers were able to add trustworthy and effective items in item bank.

Population and Sample Selection

A total 609 primary schools of district Lahore were the population of the study in which 340 are boys and 269 girls schools. Total 122,652 students are enrolled in Primary schools of district Lahore out of which 71,807 are boys students and 50,845 are girls students (School Education Department, Government of the Punjab, 2018). Accessible population of the study includes two tehsils of district Lahore which are Cantt and Shalimar. These Tehsils include total 270 primary schools from which 116 are girls and 154 are boys (School Education Department, 2018).

Table 1: *Gender wise Distribution of the Accessible Population*

Tehsil	**Schools**		**Total Schools**	**Students**		**Total**	**Students ***
	Boys	Girls		Boys	Girls		
Lahore Cantt	77	61	138	1,436	813	2,249	
Shalimar	77	55	132	1,782	1,052	2,834	
Total	154	116	270	3,218	1,865	5,083	

*Grade 5

Table 1 shows the reachable population of district Lahore that was divided into 5 tehsils but the researchers has selected 2 tehsils (Lahore Cantt and Shalimar) as per convenient. These tehsils include total 270 schools from which 154 are boys and 116 are girls. These tehsils include total 5,038 students of grade 5 from which 3,218 are boys and 1,865 are girls (School Education Department, 2018).

Multistage sampling technique was used to select sample. At first stage convenient sampling technique was used to select two tehsils (Lahore Cantt & Shalimar) out of 5 tehsils of district Lahore. At second stage, purposive sampling technique was used to select 24 primary schools (12 boys and 12 girls). According to proportion of population 18% students were selected from Public schools. At third stage, 40 students were selected randomly from each school to administer the test. Total 960 students were selected both in the subject of science and mathematics of grade 5 out of which 480 were boys and 480 were girls.

Table 2: *Gender wise Distribution of the Sample*

Tehsil	**Schools**		**Total Schools**	**Students**		**Total**	**Students ***
	Boys	Girls		Boys	Girls		
Lahore Cantt	6	6	12	240	240	480	
Shalimar	6	6	12	240	240	480	
Total	12	12	24	480	480	960	

*40 Students from each school of Grade 5

Instruments' Development and Validation

The researchers developed 300 items, 150 in the subject of Science and 150 in the subject of Mathematics that was split into two tests, each of 75 items, as students of grade 5 cannot sit long time for the test. Each test was administered on 960 students of grade 5. The contents of the tests were based on the textbooks of Science and Mathematics of grade 5. Total 90 minutes were given to students for the completion of each test and before taking the second test, the researchers given refreshment to motivate students. So, that the students feel fresh and energetic for science test. Both tests were validated by consulting one university teacher and two subject teachers of Science and Mathematics. In view of their valuable opinions both tests were improved in terms of format\ style, content and language. Pilot testing was conducted in 2 public schools of Lahore. For the pilot testing 50 students were selected. Test duration was of two hour and thirty minutes. After test was piloted, test items were reviewed and improved. Table of specification was developed for both subjects Science and Mathematics. The researchers also reviewed the grade 5 Curriculum of Science and Mathematics 2006 and the weightage of different e.g. abilities in the test of Science and Mathematics were 50% Knowledge, 20% Understanding and 30% Application.

Data Analysis and Results

Data was analyzed by using ConQuest software. ConQuest software provides the item response analysis that was used to represent results. Total 300 items were analyzed, 150 in the subject of Science and 150 in the subject of Mathematics.

Table 3: *Interpretation of Values for Item Difficulty*

Very Easy	0.91 and above
Easy	0.76 to 0.90
Optimum difficulty	0.26 to 0.75
Difficult	0.11 to 0.25
Very Difficult	0.10 and below

(Source: Ebel, 1991)

Table 3 shows the range of item difficulty. Values above 0.11 will be considered acceptable. Values below 0.10 will not be acceptable as they are showing that the question is very tough. Items are considered good when their difficulty is optimum 0.26 to 0.75.

Table 4: *Interpretation of Values for Discrimination*

High	0.40 and above
Moderate	0.20 to 0.39
Low	0.19 and below

(Source: Ebel, 1991)

Table 4 shows the range of item discrimination. Values above 0.20 will be considered acceptable. Values below 0.19 will not be acceptable as they are showing very low discrimination value. Good items have high discrimination power 0.40.

Findings of the study were discussed with respect to each research question.

Question 1. What will be the viable steps for the development of item bank in the subject of Science at grade 5?

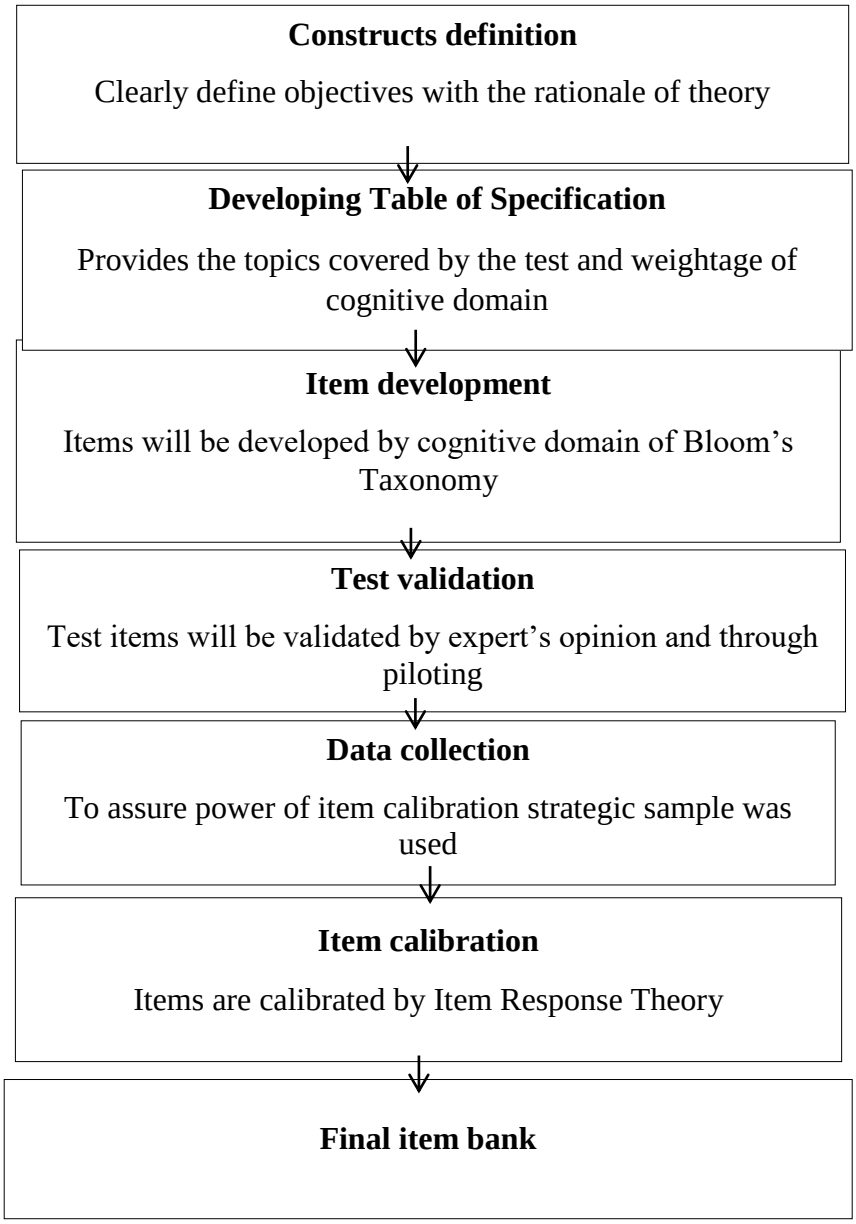

Figure 1. Steps of Item Bank Development

Question 2: How far developed items using IRT can assess lower cognitive abilities of students in the subject of science at grade 5 level?

After analyzing the developed items of science of grade 5 by using ConQuest software the researchers can clearly depicts that the developed items assess the lower cognitive abilities of the students or not. ConQuest software provides the values of item difficulty and discrimination that

are the most important information to determine whether the developed item is according to ability of the student or not.

Table of specification of grade 5 science is developed before the construction of the test. In table of specification total 8 chapters are present. These chapters are divided into three lower cognitive levels such as Knowledge, Understanding and Application according to Bloom's taxonomy. The weightage of cognitive levels are Knowledge = 50%, Understanding = 30% and Application = 10%. This weightage is provided by Punjab Examination Commission. Total 150 items are developed according to these cognitive levels. After the development of the test, it was administered on the 980 students of grade 5 then these items were analyzed by using ConQuest software. Total 58 items of science of knowledge level were accepted out of 76 items and 18 items are rejected due to its poor difficulty and discrimination value. Total 37 items of science of understanding level was accepted out of 44 items and 7 items were rejected. Total 20 items of science of application was accepted out of 30 items and 10 items were rejected. Total 113 items are added in the item bank of science that clearly depicts that these items accurately measures the abilities of the students of grade 5. Total 113 items are added in the item bank of science that clearly depicts that these items accurately measures the abilities of the students of grade 5.

Question 3: How far developed item bank of science at grade 5 is valid and reliable?

Validity and reliability of the science item bank of grade 5 is determined by the values of item difficulty and discrimination obtained from ConQuest software. Summary statistics of science item bank are given below.

Table 5: *Summary Statistics of Grade 5 Science Items*

Total No. of Items	Good Items	Fair Items	Poor Items	Eliminated Items
150	81	32	22	15

Table 5 shows that total 150 items of science were analyzed by using ConQuest software, out of which 81 items are of good quality because their difficulty level is moderate (0.26 to 0.75) and discrimination level is relatively high (> 0.40) . Item is considered fair when the discrimination value of item is optimum and difficulty level is easy, optimum and difficult .Total 32 items are fair, because their difficulty level is easy (0.76 to 0.90), optimum (0.26 to 0.75) and difficult (0.11 to 0.25). Discrimination level of fair item is moderate (0.20 to 0.39). Total 22 items are poor because their difficulty level is easy (0.76 to 0.90), optimum (0.26 to 0.75) and difficult (0.11 to 0.25). Discrimination level of poor item is below 0.19. Total 15 items were having negative discrimination that needs to be eliminated. Hence, 113 items of science are accepted and 37 items are rejected.

Question 4: How far developed items using IRT can assess lower cognitive abilities of students in the subject of mathematics at grade 5 level?

After analyzing the developed items of mathematics of grade 5 by using ConQuest software the researchers can clearly depicts that the developed items assess the lower cognitive abilities of the students or not. ConQuest software provides the values of item difficulty and discrimination that are the most important information to determine whether the developed item is according to

ability of the student or not. Table of specification of grade 5 mathematics is developed before the construction of the test. In table of specification total 9 chapters are present. These chapters are divided into three lower cognitive levels such as Knowledge, Understanding and Application according to Bloom's taxonomy. The weightage of cognitive levels are Knowledge = 50%, Understanding = 30% and Application = 20%. This weightage is provided by Punjab Examination Commission. Total 150 items are developed according to these cognitive levels. After the development of the test, it was administered on the 980 students of grade 5 then these items were analyzed by using ConQuest software. Total 55 items of mathematics of knowledge level were accepted out of 76 items and 21 items are rejected due to its poor difficulty and discrimination value. Total 34 items of mathematics of understanding level was accepted out of 44 items and 10 items were rejected. Total 22 items of mathematics of application was accepted out of 30 items and 8 items were rejected. Total 111 items are added in the item bank of mathematics that clearly depicts that these items accurately measures the abilities of the students of grade 5.

Question 5: Whether the developed item bank of mathematics at grade 5 is valid and reliable?

Validity and reliability of the mathematics item bank of grade 5 is determined by the values of item difficulty and discrimination obtained from Co*Q*uest software. Summary statistics of mathematics item bank are given below.

Table 6: *Summary Statistics of Grade 5 Mathematics Items*

Total No. of Items	Good Items	Fair Items	Poor Items	Eliminated Items
150	15	96	29	10

Table 6 shows that total 150 items of Mathematics were analyzed by using ConQuest software, out of which 15 items are of good quality because their difficulty level is moderate (0.26 to 0.75) and discrimination level is relatively high (> 0.40). Item is considered fair when the discrimination value of item is optimum and difficulty level is easy, optimum and difficult. Total 96 items are fair, because their difficulty level is easy (0.76 to 0.90), optimum (0.26 to 0.75) and difficult (0.11 to 0.25). Discrimination level of fair item is moderate (0.20 to 0.39).

Total 29 items are poor because their difficulty level is easy (0.76 to 0.90), optimum (0.26 to 0.75) and difficult (0.11 to 0.25). Discrimination level of poor item is below 0.19 that shows the items are having low discrimination index so, the items needs to be revised. Total 10 items were having negative discrimination that needs to be eliminated. Hence, 111 items of science are accepted and 39 items are rejected.

Discussion

The purpose of this study was to develop and validate item bank using item response theory at primary level. The study was quantitative in nature. The researchers developed 300 items, 150 in the subject of Science and 150 in the subject of Mathematics that was split into two tests, each of 75 items, as students of grade 5 cannot sit long time for the test. Each test was administered on 960 students of grade 5.

After collecting the data, items were analyzed by using ConQuest software. Total 300 items were analyzed 150 in the subject of Science and 150 in the subject of Mathematics. The results of item calibration of Science items revealed that 81 items are of good quality, 31 items are fair, 24 items are poor that needs to be revised and 14 items were having negative discrimination that needs to be revised or eliminated and the results of Mathematics items reveals that 15 items are good, 97 items are fair, 28 items are poor that needs to be revised and 10 items having negative discrimination power that needs to be eliminated. Total 76 items were eliminated from 300 items from the scale and 212 items were accurately falls on the criteria of the item calibration.

Burghof (2001) conducted a relevant research in Geography for accumulating an item bank for computerized linear and adaptive testing. 152 multiple-choice items were developed relating to Geography for grade 8^{th} and 9^{th} on maps and mapping skills. The data were analyzed by ConQuest software. All items were stored in the item-bank because all the items of geography on long linear scale were having difficulty value between -4 to +4. Results of all these studies indicate that item bank is the future of assessment as it develops the items according to the abilities of the students. Baumeister, (2013) conducted a research on the calibration and creation of an item bank for the assessment of activities of daily living in cardiovascular patients using Rasch analysis. I81 items were composed, and these items were answered on a five-point Likert scale by 720 cardiovascular diseases patients, who were recruited in fourteen German cardiac rehabilitation centers. Rasch analysis based on the partial credit model was conducted to test for unidimentionality and to calibrate the item bank. After Rasch analysis, 33 items were accepted and placed in an item bank. Hence, it provides the basis for the development of a CAT for the assessment of patients with cardiovascular diseases. This research reveals that item bank is not important in the educational field but it also finds its roots in health department. This research is not relevant to my subject but the process of analyzing the item was same.

Bjorner and Thissen (2007), conducted a research in computerized adaptive assessment and developing tailored instruments. The basic purpose of this research was to determine the steps required for developing an item bank or how item response theory helps in psychometric analysis that is relevant to this study's first research question that is what are the steps required in developing an item bank. According to this research, item bank development requires careful attention to construct definition, to item selection and item development, to the selection of the developmental and norming samples, and to the psychometric analyses. This research recommended that psychometric analyses should be done by the Rasch Unidimensional Measurement Model (RUMM 2010) computer program or ConQuest software. Chaowprapha and Wohler (2008), conducted a related study to develop computerized adaptive test and Mathematics item bank in Thailand for the students to cross-examine the bank. On mathematical equation total 290 multiple choice test items were composed for an item bank. Test was administered to 3,062 students of grade 6 students. Rasch Unidimensional Measurement Model (RUMM 2010) computer program was used to analyze the data. Ninety-eight test items were added in the mathematics item bank as these items were fulfilling the criteria of item calibration and. After trialing computer program for computerized adaptive testing will be improved, tested and generated.

Conclusions and Recommendations

On the basis of findings, following conclusions and recommendations were made. Out of 150 items of Science, 81 items are of good quality, 32 items are fair, 22 items are poor that needs to be revised and 15 items were having negative discrimination that needs to be eliminated. Out of

150 questions of Mathematics 15 items are good, 96 items are fair, 29 items are poor that needs to be revised and 10 items having negative discrimination power that needs to be eliminated. Total 224 items were added in item bank, 113 in the subject of Science and 111 in the subject of Mathematics. The non-performing items of the mathematics and science test (76 items out of the 300) were eliminated from the scale, leaving 224 items that fitted the measurement of item response theory.

From the results of the study it was recommended that the developed items of Science and Mathematics could be used on large scale to measure the abilities of the students. Items need to be developed according to the weightage of cognitive abilities because an item may show low discrimination if the test measures many different content areas and cognitive abilities. Teachers could prepare students of Grade 5 for Punjab Examination Commission (PEC) by using these items in the classroom. New technologies can be incorporated into assessment such as computerized item banking. Computerized item banking is an advanced technology that streamlines production, procedures, and constructs psychometrically sound tests. Avoid developing extremely difficult and easy items as such type of items have low ability to discriminate.

References

Andrich, D., Luo, G., & Sheridan, B. (2003). *RUMM 2010: A window based item analysis program employing Rasch unidimensional measurement model*. RUMM.

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: *ERIC Clearinghouse on Assessment and Evaluation.* Original work published in 1985. Retrieved from http://echo.edres.org:8080/irt/baker/

Baumeister, H. (2013). Development and calibration of an item bank for the assessment of activities of daily living in cardiovascular patients using Rasch analysis. *Health and Quality of Life Outcomes, 11(*2), 133-136.

Bjorner, J.B., & Thissen, D. (2007), Developing tailored instruments: item banking and computerized adaptive assessment. *Springer Science and Business Media, 16*(2), 95-108.

Burghof, K.L. (2001). Assembling an item bank for computerized linear and adaptive testing in geography. *International Educational Journal*, *2*(4), 101-106.

Burghof, K.L. (2001). *Computerized adaptive testing (CAT) in society and environment (SOSE)*. (Unpublished B. Ed Honors Thesis, School of Education, Flinders University, Adelaide).

Chaowpraphа, C., & Wholer, R. (2008). *Item banking with Rasch measurement*: An example for primary mathematics in Thailand. Paper presented at the international conference on sustainability in higher education: Direction for change, Perth Western Australia, November 19-21, 2008.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. Oxford University Press.

Ebel, R.L. & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall: Englewood Cliffs.

Friyatmi, Mardapi, D., Haryanto, & Rahmi, E. (2020). The development of computerized economics item banking for classroom and school-based assessment. *European Journal of Educational Research*, 9(1), 293-303. https://doi.org/10.12973/eujer.9.1.293

Grolund, N. E. (1998). *Assessment of student achievement* (6th ed.). Allyn & Bacon.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation.* Sage.

School Education Department (2018, August 18). *Importance of school education for economic growth.* Government of the Punjab. Retrieved from http://schools.punjab.gov.pk/

Stoeger, J. (2017, April 14). *Assessment system of good measure*. Retrieved from http://www.assess.com/item-banking-can-improve-assessment

Van der Linden, W. J. (1986). Computerized educational testing. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 138-150). Pergamon Press.