Journal of Science Education
Vol. 7, Issue 2, 2025, 87-96

Using Item Response Theory to Develop and Validate Item Bank in Elementary Grade Mathematics

Muazzam Dildar[1]
Nauman Ahmed Abdullah[2]

Abstract

In Pakistan, educational institutions use a variety of assessments, of which tests are the most often used type at all levels. The creation of item banks is one of the contemporary test trends. The development and validation of an item bank in mathematics for elementary-grade pupils was the aim of this study. The study's approach was quantitative, and a math test for grade 8 was created using a table of specifications. Initially, the test had 150 items. After applying item difficulty and discrimination index, 36 items were excluded due to poor values on these indices. The way these items were changed, the discrimination index and difficulty level of 18 items were both fairly good. Due to the regulated difficulty level and discrimination index, 96 items were also included. In the end, the researchers obtained 114 items for the study's target audience, which was intended to be administered to 720 Grade 8 pupils in the Lahore division. ConQuest software was used to evaluate the data, and it was discovered that 33 of the items were good, 31 were fair, 30 were put in the poor grade category, and 22 were eliminated. These resources can help teachers better prepare Grade 8 pupils for the Punjab Examination Commission. To enhance the value of assessment, item bank may be developed for all educational levels.

Keywords: Item Response Theory, Grade 8, Mathematics, Item Bank, Item Difficulty, Item Discrimination.

[1] MPhil Education, Instructor, Department of Education, Virtual University of Pakistan.

[2] Corresponding Author: Assistant Professor and In-charge,
Syed Babar Ali Department of Education, GC University Lahore.
Emails: nauman.abdullah@gcu.edu.pk, nauman101@hotmail.com
ORCID: 0000-0003-4435-5674

Introduction

One of the essential elements of a successful education sector at all educational levels is the requirement for quality assurance. The central query is: how do we evaluate quality? The most popular technology and method used by teachers and institutions to support quality assurance is the "item bank." To measure a predetermined set of concepts, a purposeful store of stuff is created through item banking (Stoeger, 2017). An organized, categorized, and collated set of test objects is referred to as an item bank (Choppin, 1985).

Quality evaluation in the modern era is dependent on item banks. Item banks are collections of several appropriate test items arranged according to focus area, educational level, determined instructional goal, and certain connected item characteristics such as item complexity and discriminating power (Gronlund, 1998, p. 130).

Articles with psychometric qualities are available on item banks. Two psychometric techniques exist. The prevalent approach is classical test theory, which emphasizes test-retest reliability, internal consistency, various types of validity, standardization, and normative data. The focus of current test theory, also known as item response theory (IRT), is on how various test objects function in construct evaluation. IRT permits the construction of parallel test forms, the size of test items for difficulty, and the use of adaptive computer testing (DeMars, 2010). Instead of using a composite of the item outcomes as a test mark, the key components of the item response principle are concentrated on the individual test items (Baker, 2001). Because IRT systems are frequently far more traditional and user-unfriendly than other commercially available statistical bundles, using IRT methods for data analysis can be challenging (Kline, 2005).

George Rasch created the IRT model in Denmark in 1960 as a means of assessing reading proficiency and creating examinations for the armed forces. One of the best IRT models uses his work. Four chapters of IRT were added by Allan Birnbaum to Lord and Novicks' 1968 statistical theories. Lord worked hard for the "Educational Testing Service (ETS)," and as a result, he was given excellent access to the database. Gradated Response Models were introduced by Samejima in the IRT in 1969, and she forever changed the IRT models that are connected to Likert-scale data and other MCQ tests (DeMars, 2010).

IRT Theory is used to calibrate the questions since they provide more expected statistical information and are therefore preferred to be included in the item bank to calculate the students' cognitive abilities and the likelihood that they would provide accurate responses. It compares the difficulty of questions using an IRT model. IRT models' primary advantage is that test and question features do not fluctuate, in contrast to classical test theory, which allows both item and test characteristics to change (Hathaway et al., 1985). A calibrated IRT model, such as the 3PLMor Rasch model, enables the identification of questions with poor performance as well as the equating of items onto a single scale. While the Rasch model calibration will require 100 to 200 students (Linacre, 1994), the 3PLM parameter analysis will require 1000 to 2000 students (Green et al.,1984).

A study on the creation and validation of a math accomplishment test was done by Jayanthi (2014). The study involved pupils in the 10th grade and was carried out in the Chennai district. The 10th standard Mathematics curriculum was used to create a 150-item math achievement test. 327 pupils from five schools in that district made up the test's sample. After review and assessment, grades were given. The performed items analysis sheets were used to calculate the discrimination and difficulty indices.

Significance of the Study

There is a lot of work to establish this model in Pakistan, where item banking is in a developing stage. The grade 8 mathematics item banks will receive better information as a result of this analysis. Teachers could utilize this study to construct tests faster and with less wasted time. This item bank is free for teachers to use. It might reveal the items' difficulty and discrimination indices, which would ensure the accuracy of the assessment. The Punjab Examination Commission (PEC) will find this study extremely helpful for creating math tests and assignments for grade 8. This study will serve as a guide for aspiring teachers on how to create test objects.

This study might help develop a revised curriculum that promotes the value of assessment and improves teachers' ability to memorize and recall test topics. It would give a better knowledge of the settings and results of grade 8 mathematical studies. It also plays a crucial role in determining the value of evaluation that improves students' learning capacities.

Objectives of the Study

The study was designed around the following objectives:

1. To create an item bank for the grade 8 math course.
2. Using item response theory, validate grade 8 math item bank.

Research Methodology

The positivist philosophical research paradigm served as the basis for this study's analysis of the social world. Researchers who focus on positivism-based pedagogy frequently employ quantitative techniques to collect quantifiable numerical data (Charles & Bawa, 2017). The 8th grade students who participated in this quantitative study took a pilot test on the topic of mathematics. The researchers created and validated 150 items using item response theory from the mathematics textbook for the eighth grade. The test had about 150 multiple-choice questions. The psychometric characteristics of each item were tuned. The researcher employed a multistage sampling strategy to gather data from several students. Following the analysis, the researcher incorporated trustworthy and valid MCQs to the item repository.

Population of the Study

The research's target area includes all male and female students in the public sector enrolled in elementary schools in the Lahore division. In Lahore division, there were a total of 884 elementary schools, of which 401 were for boys and 487 were for females. There were a total of 124,703 children enrolled in the district of Lahore's elementary schools, of which 64,923 were female and 59,780 were male. (Department of School Education, 2018).

The study's sample was chosen using a multistage stratified random sampling process. Only two districts from the Lahore Division were randomly selected for the first stage, which were Lahore district and Sheikhupura district. The second stage was the random selection of two tehsils from each district. Two selected districts were Lahore Cantt and Model Town from Lahore district and Shiekhupura and Ferozwala from Sheikhupura district. These tehsils had a data of 7240 students. The Stratified Random Sampling Technique was used to select 6 boys and 6 girls from elementary schools in each tehsil in the following stage. At the end, 30 pupils from each school were conveniently taken. The final sample size selected was 720 pupils in total, including 360 girls and 360 boys, such sample size is acceptable is item validation studies (Friyatmi et al. 2020; Shahid et al. 2023).

Instrumentation

The researchers created three tests, each with 50 items, totaling 150 items. Holman, et al. (2003) developed 75 items and validated those through IRT, similarly, Shahid et al. (2023) developed 150 items for math exam and then split the items into two sets of 75. This establishes the appropriateness of the number of items developed and validated using IRT in this study. The tests' subject matter was drawn from mathematics for class 8. The pupils were given one hour to complete each test, which had 50 multiple-choice questions. The researchers also studied the mathematics curriculum document for the year 2006 again, which helped them comprehend the levels of Bloom's Taxonomy that are crucial for elementary school kids. The opinions of the specialists present at the Institute of Education and Research (IER), University of the Punjab, Lahore, validated the validity of the test.

Two technical experts from the department of assessment and research evaluation at IER faculty, and one topic expert who teaches mathematics at science education department at IER were consulted for the validation of item bank. Additionally, assessments were piloted on 40 eighth-grade students to identify item difficulty and item discrimination. The difficulty and discrimination index of each item, along with explanations for them, were included in the item analysis. The pilot testing consisted of 150 MCQs in total. 38 multiple-choice questions had poor item difficulty and discrimination, which is why these items were eliminated. The way these things were changed, the discrimination index and difficulty level of 18 items were fairly good. Due to regulated difficulty level and discriminating index, 96 items were also included resulting in a final 114 items.

Analysis and Finding

ConQuest Software was used to assist in the data analysis. This programme analyses the data that was used to get the results. It was decided to analyse 114 questions from the Mathematics curriculum. The analysis of grade 8 math assessment on the data of 720 pupils is presented below.

Table 1
Range of Item Difficulty

Item Difficulty Range	Difficulty Level
< 0.91	Very Easy
76%-90%	Easy
75%-26%	Average
25%-11%	Difficult
>10%	Very Difficult

(Source: Ebel & Frisbie, 1991)

The value range of the difficulty (p) levels for the chosen items is shown in Table 1 above. Values greater than 0.91 were interpreted as very easy. Values less than 0.10 were seen as unsatisfactory and unacceptable because they indicate that the questions are significantly more challenging. When the difficulty value was considered optimal, between 26% and 75%, items are reflected as great items for a normal curve assessment.

Table 2
Range of Discrimination Power

Discrimination level	Value Range
Higher	0.41 >
Normal	0.18 to 0.41
Below average	0.18 <

(Source: Ebel & Frisbie, 1991)

The value range of an item's discrimination (D) is displayed in Table 2 above. Values greater than 0.18 were considered to be acceptable. And values that are less than 0.18 will be disregarded since they demonstrate a noticeably low level of discrimination. Good things have a high power of discrimination, which is 0.41.

Table 3

Statistical Summary of Class 8 Mathematics items

Total items	Excellent items	Avg. items	Below Avg. items	Excluded items
114	33	31	28	22

The ConQuest Software was used to assess 114 test questions for the grade 8 subject of mathematics. Table 3 highlights that of them, 33 test items produced outstanding results because their difficulty ranges were average (28% to 74%) and their discrimination powers were high (> 0.45). These things are regarded as good items since they have an average level of discrimination and a reasonable range of simple difficulty. 31 test items produced an average result since the average item's discrimination power falls within the acceptable range of 0.19 to 0.37 and the range of difficulties includes simple (74%–88%), ideal (24%–74%), and challenging (12%–24%).

The difficulty levels of the next 28 items range from easy (0.76 to 0.90), optimal (0.26 to 0.75), and difficult (0.26 to 0.75). (0.11 to 0.25). Poor item discrimination level is less than 0.19. However, the number of items with negative discrimination that needed to be eliminated was just 22. As a result, 64 test items from the eighth-grade mathematics exam were allowed, whereas 50 test items were ignored. The researcher stated that the items used in the research may easily assess the cognitive capacities of the students by using the Conquest software for test items of Mathematics for grade 8. The ConQuest software provided the necessary data for this purpose in order to guarantee the validity of items for evaluating the student's cognitive ability.

Prior to the finalization of the exam structure and material, a table of specifications for class 8 was created. The table of specifications includes a total of 10 chapters. The Bloom's Taxonomy places these chapters in the lower cognitive stages. Knowledge is weighted at 20%, followed by understanding and application at 40%. Punjab Examination Commission is responsible for assigning this weight.

These 114 multiple-choice questions (MCQs) are aligned with Bloom's degrees of cognition. When the math test for class 8 was created, it was given to 720 students. These things were then carefully examined using conquest software. Only 20 of the 26 Mathematics exam items for knowledge level were selected, and six test items were dropped owing to their poor quality. Only 23 of the 40 math exam items for the application level were chosen, while 17 test items were left off. Following this procedure, 64 items were gathered and added to the Mathematics item bank to improve its accuracy and maintain its reliability.

Conclusion and Recommendations

It is challenging to create dependable and valid things. This study's setting makes it clear that an item bank was created for the subject of mathematics. Out of 150 test items, 31 were fair, 33 were of acceptable quality, 28 were below average and needed to be evaluated, and 22 were negative discrimination questions that needed to be excluded. But in the end, the Mathematics item bank could only manage 64 test items. The math test items that performed below average were removed from the scale, leaving 64 test items that complied with Item Response Theory.

The following are some suggestions based on this study:

1. The designed test items may be useful at every stage for assessing the students' mathematical proficiency.
2. The items used to assess students' aptitudes must consider their cognitive skills, as they can yield findings that are difficult to differentiate between across domains and cognitive abilities.
3. Examining bodies must employ the IRT model while developing tests.
4. Steer clear of making items that are overly simple or complex because they won't yield accurate discrimination indices.
5. When preparing class 8 children for the PEC board examination, teachers should use these exam items.
6. Item banking should be created for all educational levels, including primary, secondary, and higher education, to support the quality evaluation.
7. Item banking is a cutting-edge method for creating and creating psychometrically sound tests. Because of this, it can, in my opinion, be included in the evaluation.
8. To test the test items, additional primary courses should also employ item banks.

References

Baker, F. B. (2001). *The basics of item response theory*. *ERIC clearinghouse on assessment and evaluation.* Original work published in 1985. Retrieved from http://echo.edres.org:8080/irt/baker/

Charles, K., & Bawa, A., (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education, 6*(5), 26-41.

Choppin, B. (1985). Principles of item banking. *Evaluation in education: An International Review Series, 9*(1), 87-90.

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.

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.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

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

Hathaway, W., Houser, R., & Kingsbury, G. (1985). *A regional and local item response theory-based test item bank system.* (ERIC Document Reproduction No.ED 284 883).

Holman, R., Lindeboom, R., Glas, C. A., Vermeulen, M., & Haan, R. D. (2003). Constructing an item bank using item response theory: The AMC linear disability score project. *Health Services and Outcomes Research Methodology*, *4*, 19-33. 10.1023/A:1025824810390

Jayanthi, J. (2014). Development and validation of an achievement test in mathematics. *International Journal of Mathematics and Statistics Invention, 2*(4), 40-46.

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

Shahid, M., Saeed, S., & Akhtar, M. (2023). Developing and validating item bank in science and mathematics at primary level using item response theory. *Journal of Interdisciplinary Educational Studies, 3*(2), 47-57.

Stoeger, J. (2017, April 14). *Assessment system of good measure*. Retrieved from

http://www.assess.com/item-banking-can-improve-assessment.

Citation of this Article:

Dildar, M., & Abdullah, N.A. (2025). Using item response theory to develop and validate item bank in elementary grade Mathematics. *Journal of Science Education,7(2*), 87-96.