Full text open access online (Since 2009)
© KRE Kamla-Raj IJES 2026
PRINT: ISSN 0975-1122 ONLINE: ISSN 2456-6322
Int J Edu Sci, 53(2): 64-75 (2026)
DOI: 10.31901/24566322.2026/53.2.1443

A Critical Analysis of Large-Scale Math Assessments of Grade 4: Consequences for Development and Educational Optimisation

Aisha Bibi[*] and Mushtaq Ahmad

Department of Educational Sciences, National University of Modern Languages (NUML), Islamabad, Pakistan

KEYWORDS Item Difficulty Indices. Large Scale Assessment. Math Achievement

ABSTRACT The aim of this work was to portray and compare the learning achievement of Grade 4 students in mathematics subjects for improvements in curriculum developments and improved practices. The research employed a norm-referenced assessment approach to compare performance trends across different populations and geographical areas by concentrating on relative outcomes. Data of 10,578 Grade 4 students from 633 schools was critically analysed. Findings revealed that students had lower scores in Statistics and Probability items across all content areas of mathematics. Additionally, the highest score was observed in each of the knowing and reasoning domains. Students' average score in the applying and reasoning domain suggests that although students may comprehend mathematical concepts, they have difficulty applying them to solve problems. It emphasises the importance of focused teaching methods to improve students' ability to effectively utilise scientific concepts. On the basis of these findings, several implications were also outlined.

INTRODUCTION

Curriculum reforms are highly important in the arenas of science, mathematics and STEM based education to address the evolving needs of students, the workforce, and society by providing a complete and interdisciplinary education. All these reforms focus on students' creativity, critical thinking, and problem-solving abilities to solve real-world tasks (Nurtamam et al. 2023; Rahman et al. 2021). In addition to curriculum reforms, the implementation of these reforms is also challenging in the dynamic and quickly changing modern world that extends beyond traditional academic limits, and advancement in society.

Multiple techniques were employed to determine and regulate the implementation of curriculum reforms, including monitoring students' outcomes and achievements, teacher training and support, and feedback from different stockholders. Students' outcomes and achievements are primarily judged by academic assessments (for example, exams, assignments and projects) and teacher observations. However, to evaluate students' knowledge and skills across different populations, schools and geographical areas, large-scale assessments or standardised testing are frequently used (Kolomuc and Demir Saglam 2025; Morris et al. 2025; Yang et al. 2024). Standardised testing also ensures that students acquire content knowledge as well as essential competence by aligning their learning objectives with workplace demands. This promotes extensive preparation for future career challenges. Besides, teachers can use test results to assess whether students are satisfying standards and benchmarks, allowing for a more flexible approach to instruction that adapts to the needs of each individual student. Therefore, the current study has focused on this avenue of standardised testing for the implementation curriculum reforms.

In Pakistan, the National Achievement Test (NAT) serves as a standardised testing tool for primary and secondary levels of education. NAT is a sample-based assessment that takes place approximately every two years. Recently, it was conducted on 18-19 May 2023 by the National Assessment Wing, Ministry of Federal Education and Professional Training (Pakistan Institute of Education 2024). These assessments include renowned international studies that provide education leaders insight into how their nation's educational performance ranks other countries, resulting in substantial modifications to policies (Hopfenbeck 2016; Meyer and Benavot 2013). Furthermore, recently held NAT includes

[*]*Address for correspondence:*
Dr. Aisha Bibi,
Department of Education Sciences,
National University of Modern Languages (NUML),
Islamabad, Pakistan,
E-mail: aishabibi@numl.edu.pk

different subjects and domains (Math, Science, Urdu, English and Sindi) to assess the implementation of the revised 2022 National Curriculum. The scope of current studies is the evaluation of test results of Math (for Grade 4 only). Since the test's quality is also important, the NAT assessment tool was also critically analysed using item analysis, the difficulty index (DI), the discrimination index (DI), and the reliability assessment for improved practices (Kohler et al. 2020).

Content and Cognitive Areas of Math and Science Tests

In Pakistan, the Ministry of Federal Education and Professional Training has implemented the 2022 National Curriculum for Grades 1 through 8, which serves as the basis for the mathematics and science framework (Pakistan Institute of Education 2024). This framework is not meant to serve as a teaching and learning manual, rather, it outlines the goal, layout, content, and cognitive areas of the math tests. Additionally, it is essential to simultaneously focus on students' cognitive domains and skills in order for subject knowledge to be effectively integrated and applied to a variety of situations. To further align Pakistan's learning assessments with international norms, the cognitive domains have been made consistent with the TIMSS (Trends in International Mathematics and Science Study) assessment framework under the national curriculum. The national curriculum lists the following cognitive domains. In knowing the domain, students are required to be familiar with terminology and symbols and comprehend the fundamental concepts that explain them. In applying domain, students ought to be able to select and employ proper mathematical concepts and techniques in this area to solve problems in the real world. Pure mathematical questions are covered, such as equations, geometric figures, statistical data sets, and algebraic or numeric expressions. Additionally, in the reasoning domain students must apply their past mathematical knowledge in unfamiliar situations in this domain. In order to solve real-life problems, it analyses, synthesises, and evaluates data while taking into account when there is enough reliable information. Overall, the focus of this work is to critically analyse the valuable data collected by the Ministry of Federal Education and Professional Training through active support of the Pakistan Institute of Education to promote science and mathematics education in the country.

Research Objectives

Following objectives were established for the study:

1. To evaluate the assessment tool on the basis of appropriateness of used items, their difficulty level and discriminant indices. This objective was designed to improve the quality of math assessments by making sure that the items effectively differentiate between various levels of students' understanding.
2. To examine and compare the students' marks in mathematics in three cognitive domains as specified by educational standards (for example, Bloom's Taxonomy) across different provinces. This comparative analysis will provide practical insights that may result in targeted measures and modifications to math and science education in different regions.
3. To examine and compare the students' marks in mathematics on the basis of different content domains. The purpose of this objective is to identify the areas of strength and where their comprehension needs to be strengthened.

Theoretical Framework

The Classical Test Theory (CTT) has been an extensively utilised technique to assess the reliability and various attributes of measuring tools. The Classical Test Theory (CTT) is appropriate for small-scale assessments or groups with minimal psychometric knowledge because it provides an accessible approach for item analysis. It provides three main aspects of the test score (also called the observed score), the real score, and the error score (sometimes called random measurement error). Classic Test Analysis (CTA) utilises classic item-specific and sample-related statistics, such as item difficulty evaluations, item discrimination estimates, distractors' analyses, and related measures (Pakistan Institute of Education 2024).

The CTT is the theoretical framework used for this investigation which employs two primary statistics, and the item difficulty index assesses whether an item was too simple or too difficult. The DIF's range is 0-1, and the p-value is converted to a percentage by multiplying it by 100. An easy inquiry, object, or task is indicated by a high calculated value. Items with a p-value of less than 20 percent are deemed too difficult, and items with a p-value of more than 90 percent are deemed too easy, while extreme outliers are unacceptable. According to Mukherjee and Lahiri (2015), items with a DIF of more than 70 percent are too easy, those with a DIF of 30-70 percent are average, and those with a DIF of 50-60 percent are good. An indicator of whether an item discriminated between students who were knowledgeable about the subject matter and those who were not is the Discrimination Index (DI). To determine how well an item separates high-achieving children from low-achieving ones, the DI is utilised. The test-takers are split into quartiles for the purpose of computing the DI. Students with the greatest scores are in the upper quartile, those with the lowest scores are in the lower quartile, and those with average scores are in the middle quartile. The range of the value is -1 to +1. The outcome is deemed discriminatory and effective if it has higher value (Musa et al. 2018). According to Mukherjee and Lahiri (2015), if all test-takers in the higher quartile respond to the item correctly and none in the lower quartile, the DI value is 1.00. If, however, the lower group answered it properly while the higher group did not, the DI value would be -1 (D'Sa and Visbal-Dionaldo 2017). DI between 0.21 and 0.24 are acceptable, with a value of e”0.36 are excellent, items between 0.25 and 0.35 are good and items with a value of d”0.20 are poor, as classified by (Mahjabeen et al. 2017).

Research Aim

The aim of this work was to portray the learning achievement of Grade 4 students in mathematics subjects after the implementation of the recently modified curriculum (2022). The second aim was to enable the bridging of breaches by the use of assessment tools for improvements in curriculum developments and improved practices.

METHODOLOGY

For this research study, the pragmatist research paradigm was used, which emphasises realistic solutions and real-world outcomes, highlighting how well items operate within real-world situations. For item analysis, an observational research approach was employed. It provides ongoing modifications and improvements to items based on actual input from user interactions. Surveys and testing may overlook usability problems or misunderstandings that observational data might show by revealing how people interact with items. Based on data from NAT administrations in 2023, this study uses an ex-post facto approach to critically examine the development and use of NAT without the involvement of the researcher. A non-experimental research technique called ex post facto design was used to assess how well the NAT (National Assessment Test) assessment tools evaluate the desired criteria. Additionally, the Difficulty Index (DIF) and Discrimination Index (DI), two important psychometric characteristics of the NAT, were evaluated quantitatively using Classical Test Theory (CTT), which offers insights into test validity and reliability across diverse contexts.

Assessment Approach

The research employed a norm-referenced assessment approach, which is particularly effective for large-scale tests such as the National Assessment Test (NAT). This approach offers a comprehensive comparative analysis of educational accomplishment by ranking students, schools, or regions (Kolawole 2005). Without being restricted by predefined criteria, the researchers were able to investigate performance trends across different populations and geographical areas by concentrating on relative outcomes.

Population

A total of 10,578 Grade 4 students participated in the study (633 schools). The study employed scientific sampling techniques in order to ensure that the final sample of Grade 4 students was representative at the provincial and national levels in terms of gender, school type, and rural or urban location. Within each province, the sample was specifically divided on the basis of the gender of the students and whether the school was in an urban or rural location (Pakistan Institute of Education 2024). Furthermore, in order to guarantee that the final sample was representative of the student body, school sampling was conducted using probability proportional to size (PPS) for each school. Lastly, 20 individuals were selected from each of these classes' attendance records using systematic random sampling.

Research Instrument

Table 2 illustrates the content and cognitive domains of the National Assessment Test (NAT). It is essential to simultaneously focus on students' cognitive domains and skills in order for content knowledge to be effectively integrated and applied to an array of settings. 20 percent of the NAT maths test's items involved reasoning, 40 percent emphasised on application, and 40 percent evaluated knowledge (Table 1).

Ethical Considerations

The data was collected from targeted participants (Pakistan Institute of Education 2024). Related authorities and organisations were requested to give their consents for this work. This study was well aligned with university approved code (1568696934/R. Ethics/NUML).

RESULTS

The first objective of current study was to examine how well the NAT tool was developed. Keeping the study scope, the 2023 tool was subjected to item analysis in order to ascertain their psychometric qualities, particularly emphasis on objective items.

Test Reliability

The internal consistency of a math subject is measured by reliability estimates, and the measurement error of the students' unknown true score is quantified by the SEMs. According to Cronbach's alpha, the internal consistency reliability for the math subject test was acceptable, or more than 0.92, as shown in the Table 1 below. Following an analysis of the pilot items' results, the items with the best psychometric qualities for the LSA are chosen. The selection of items is based on examination of item difficulty, evaluating item discrimination and analysis of distractors.

Psychometric Analysis of NAT Test (Classical Test Theory)

The test items' reliability is typically assessed through their item difficulty and discrimination indices. Item difficulty is the percentage of candidates who properly answer a question, whereas, the rate at which the probability of successfully answering an item fluctuates in relation to an examinee's capability is known as item discrimination. According to AIR's methodology, a question was deemed too difficult if less than 20 percent of students correctly answered it. If more than 80 percent of students answered the question correctly, they were categorised as extremely easy. If the item discrimination was less than 0.20, it was considered difficult (Boateng et al. 2018). The number of items in each item difficulty category for each test is displayed in Table 2. Few items were deemed too difficult (that is, $x \leq 0.2$). Discriminations on eleven additional items, especially in Grade 4 Math, were likewise deemed to be unacceptable.

Item Difficulty Index and Discrimination of Mathematics Tests

The difficulty index of test items in mathematics is shown in Table 2. One item was deemed too challenging (that is, $x \leq 0.2$). As a result, this item should be changed or removed from the questionnaire. One test item, M 36 (Statistics, basics), falls within the difficult index range's difficulty standards. These items are either relatively sim-

Table 1: Specification for Mathematics item details with regard to content and Cognitive Domain

Subject	*Cognitive contents*	*Knowing (40%)*	*Applying (40%)*	*Reasoning (20%)*	*Total*
Math	Numbers and algebra (58%)	11 items	11 items	6	28
	Measurement and geometry (27%)	6	5	2	13
	Statistics and Probability (15%)	2	3	2	7
	Total Items	**19**	**19**	**10**	**48**

Table 2: Item Difficulty Index and Discrimination of Mathematics Tests (n = 10578)

Item No.	*Students Learning Outcomes (SLO)*	*Difficult Index*	*Discrimination*
M1	Determine and differentiate between parallel and non-parallel lines	0.79	0.31
M2	Interpret and display the time from analog and digital clocks in 12- and 24-hour formats	0.64	0.34
M3	Determine the rules of divisibility for 2, 3, 5, and 10 and apply them to four-digit	0.59	0.45
M4	Transform length measurements between different units (kilometre, metre and centimetre)	0.53	0.44
M5	Solve addition and subtraction real-life scenarios word problems	0.67	0.43
M6	Determine each digit's place value in five-digit integers	0.64	0.46
M7	Explain the result of a basic probability experiment involving a spinner and dice by employing mathematical terms (that is, less likely, more likely, equally likely, unlikely, and certain)	0.57	0.36
M8	Transform bigger capacity units (millilitres and litres) into smaller ones (litres and kilolitres)	0.33	0.48
M9	Use the characters '<', '>' or '=' to compare two numbers up to five digits	0.62	0.42
M10	Create, read, and interpret single and double bar graphs in both horizontal and vertical orientations using real-world scenarios	0.31	0.38
M11	Apply fractions to represent decimal numbers up to three decimal places	0.55	0.44
M12	Identify and distinguish between obtuse, right, and acute angle	0.50	0.46
M13	Multiply and divide mixed numbers, proper fractions, and improper fractions by a whole number	0.55	0.41
M14	Practicing rounding off figures to the nearest tens, hundreds, and thousands	0.47	0.50
M15	Determine and write expressions sentences that illustrate unknown-problem scenarios	0.52	0.42
M16	Recognise and distinguish between factors and multiples, and locate all factor pairs of a number that are common factors of two numbers	0.62	0.52
M17	Determine and distinguish between prime and composite numbers up to 50	0.50	0.48
M18	Convert mass units from kilograms to grams	0.52	0.50
M19	Fill in the patterns of increasing and decreasing numbers	0.57	0.45
M20	Find addition and subtraction word problems from everyday life	0.64	0.50
M21	Add and subtract up to five-digit integers (both with and without regrouping)	0.69	0.47
M22	Distinguish between mixed numbers, improper fractions, and proper fractions	0.38	0.49
M23	Divide and multiply mixed numbers, proper fractions, and improper fractions by a whole number	0.47	0.51
M24	Create, read and analyse single and double bar graphs related to daily life scenario in both horizontal and vertical orientations	0.42	0.41
M25	Solve multiplication real world word problems	0.38	0.52
M26	Explain a circle's diameter, radius and circumference	0.55	0.42
M27	Recall roundoff numbers	0.20	**-0.03**
M28	Divide and multiply a 2-digit number by either a 1-digit number or a 2-digit number up to one decimal place	0.57	0.53
M29	Determine the duration of various events by utilising the start and end times	0.43	0.35
M30	Fill in the provided sequences of ascending and decreasing numbers	0.46	0.25
M31	Recognise and differentiate between factors and multiples, and locate all factor pairs of a number that are common factors of two numbers	0.61	0.46
M32	Identify the place value of decimal values up to three decimal places	0.24	0.13
M33	Solve real word problems by converting, adding, and subtracting mass	0.24	0.20
M34	Divide and multiply a two-digit value by 10 and 100, with one decimal place	0.43	0.48
M35	Recognise and distinguish between obtuse, right and acute angles	0.33	0.25
M36	Using mathematical terminology, explain the results of a basic probability experiment involving a spinner and dice	**0.16**	**-0.03**
M37	Sort mixed numbers, improper fractions and proper fractions	0.55	0.52
M38	Employ formulas to determine the area and perimeter of rectilinear shapes, such as squares and rectangles	0.44	0.26
M39	Transform larger time units (hours, minutes, seconds, years, months, weeks and days) into smaller ones	0.54	0.50
M40	Contrast and arrange unlike fractions	0.53	0.31
M41	Determine each digit's place value in 5-digit integers	0.48	0.46
M42	Create, read and analyse single and double bar graphs related to daily life scenario in both horizontal and vertical orientations	0.53	0.56
M43	Find addition and subtraction word problems from everyday life	0.60	0.48
M44	Recognise and distinguish between factors and multiples, and locate all factor pairs of a number that are common factors of two numbers	0.47	0.45
M45	Identify and illustrate symmetrical lines in two-dimensional shapes and construct s ymmetrical figures relative to a specified line of symmetry	0.39	0.39
M46	Create, read and analyse single and double bar graphs related to daily life scenario in both horizontal and vertical orientations	0.41	0.52
M47	Convert decimal integers to fractions up to three decimal places	0.47	0.49
M48	Create, read and analyse single and double bar graphs related to daily life scenario in both horizontal and vertical orientations	0.60	0.53

ple or slightly difficult. However, 25 questions fall into the moderately difficult or rather simple difficulty index range (0.30-0.70), and the majority of the respondents provided the answers. Similarly, three question items of the mathematics tests should be revised or removed from the list of questions since, two worst/defective items and one marginal item among them, are marginal items but not discriminant. 33 items are considered to be moderately discriminant or fair. There are then 12 items that are either excellent or discriminant. Furthermore, these items are excellent discriminants or items under the criteria of $\geq$0.80.

Key Findings

1 item (from Probability) and 22 items were the 'most difficult' and 'difficult' items respectively, in Grade 4 Math, indicating that these items ought to be revisited, revised or removed from the test. Regarding discrimination, two additional items (domain: probability and real numbers), mainly in Grade 4 Math, were likewise deemed to be marginal and non-discriminant items and are unacceptable. Additionally, it has been found that item no M36 was an extremely difficult item question and is also a marginal discriminatory item. Sahoo and Singh (2017) studied the relationship between difficulty and discrimination indices and observed that item discrimination decreased for very simple and very difficult items but increased with increasing difficulty level. Sim and Rasiah (2006) observed that they are more likely to exhibit negative discrimination, and that these questions typically have excellent discriminating power. The number of items in each category of classical discrimination is summarised in Table 3. Two Grade 4 Math items specifically had negative discriminating indices. This indicates that students usually misinterpret how to address these specific issues.

In NAT 23, three Bloom's Taxonomy-aligned domains, that is, knowing, applying and reasoning, were used to evaluate Grade 4 math. In mathematics, students' overall average score in "Knowing" was 50 percent. Moreover, students scored an average score of 48 percent in the "reasoning" area and 48 percent in the "applying" domain as shown in Table 4. The average scores for all three skills (~48.6-50.5 percent) are similar, indicating that students' performance in knowing, applying and reasoning was comparable. The scores indicate moderate proficiency or understanding in these areas, as they fall around the middle of a normal 0-100 percent scale.

Table 4: Comparisons of the average percentage of marks achieved by students

Subject	*Cognitive domain*	*Percentage of marks achieved (Indicative standard error)*
Maths	Knowing	50.6 (3.8)
	Applying	48.6 (2.2)
	Reasoning	48.6 (4.0)

Table 5 shows comparisons of the percentage marks achieved by students across different provinces. In mathematics, interestingly, Punjab has the highest percentage at 63 percent and 64 percent in knowing and applying respectively, whereas, GB has the lowest percentage of 40 percent and 35 percent respectively, in these domains among all provinces. Regarding reasoning domain, results showed that Punjab has the highest per-

Table 3: Item evaluation and comparison of Mathematics tests based on the Difficulty and Discrimination Index

Difficulty Index Subject	Very difficult 0 $\leq$x$\leq$0.2	Difficult 0.2 $\leq$x$\leq$0.5	Moderately difficult 0.5$\leq$x$\leq$0.8	Easy x >0.8		Total
Mathematics	1	22	25	0	48	
Discrimination Index Subject	x<0 Worst/ Defective item	0 $\leq$x$\leq$0.2 Marginal item, not discriminant	0.2 $\leq$x$\leq$0.5 Moderately discriminant/ Fair	0.5$\leq$x$\leq$0.8 Discriminant, good	x $\geq$0.8 Very discriminating, very good	Total
Mathematics	2	1	33	12	0	48

Table 5: Comparisons of the standard error percentage of math marks achieved by students across different provinces (Cognitive domain wise)

		Percentage of possible marks achieved						
Subject	*Content domain*	*Balochistan*	*KP*	*Punjab*	*Sindh*	*AJK*	*GB*	*ICT*
Maths	Knowing	43.3	42.0	63.2	50.5	44.1	**40.1**	42.1
	Applying	38.8	37.3	64.6	49.0	43.2	**35.9**	36.0
	Reasoning	37.9	39.2	62.9	48.2	42.7	40.8	**36.6**

Note: GB: Gilgit Baltistan, KP: Khyber Pakhtunkhwa, AJK: Kashmir, ICT: Islamabad Capital Territory
*Indicative standard error indicates the precision of

centage at 62 percent, whereas ICT has the lowest percentage of 36 percent among all provinces.

Key Findings

Students' average scores in the applying and reasoning domain suggest that although students may comprehend mathematical concepts, they have difficulty applying them to solve problems. This can be the result of inadequate strategies for problem-solving or a lack of practice. The "reasoning" domain score of 48 percent indicates that students get stuck to draw inferences and justify their answers. This challenge could result from inadequate critical thinking training or a lack of practice with reasoning-based tasks.

The highest scores of 63 percent and 64 percent were observed in each of the knowing and reasoning domains, respectively. This shows that students can solve reasoning problems and appear to have a strong grasp of basic concepts still more practice is required. With the "applying" domain's lowest score of 49 percent, students are apparently unable to apply their knowledge in real-world situations. It emphasises the importance of focused teaching methods to improve students' ability to effectively utilise scientific concepts.

Punjab consistently scores highest on "knowing" and "applying" domains with the highest percentages. On the other hand, GB has the lowest scores in two domains of "knowing" (40% and 35%). Moreover, ICT showed the lowest value (36%) in reasoning.

Comparisons of Students' Achievement on the Basis of Content Domain

Across all content areas of mathematics, Statistics and Probability had the lowest level of performance, with students answering only 43 percent of items correctly on average (Table 6). The fact that this is a new subject that was included in the 2022 National Curriculum may have something to do with this. In addition, students' performance in algebra, measurement and geometry lagged slightly behind that of other Grade 4 math curriculum areas except probability and statistics. Consequently, the unfamiliarity may result in difficulties with understanding and successfully implementing concepts (Table 7).

Table 6: Comparisons of the average percentage of math marks achieved by students

Content domain	*Percentage of marks achieved(Indicative standard error)*
Algebra, Measurement and Geometry	48.6 (3.3)
Numbers and Operations	51.7 (2.4)
Statistics and Probability	43.0 (5.9)

Table 7: Percentage of marks achieved along with indicative standard error on the basis of Content Domain

Content domain	*Balochistan*	*KP & NMD*	*Punjab*	*Sindh*	*AJK*	*GB*	*ICT*
Algebra, Measurement andGeometry	39.7(3.3)	40.0(3.2)	**61.5**(3.7)	48.9(3.9)	43.3(2.9)	**38.4(3.6)**	**38.2(4.3)**
Numbers andOperations	42.3(2.4)	41.0(2.0)	**67.0**(3.2)	51.5(2.7)	45.3(2.2)	**40.0(2.3)**	**40.4(2.7)**
Statistics andProbability	34.7(5.6)	33.2(4.2)	**56.8**(7.9)	43.1(7.7)	37.4(5.4)	**33.3(4.4)**	**32.5(5.1)**

Note: GB: Gilgit Baltistan, KP: Khyber Pakhtunkhwa, AJK: Kashmir, ICT: Islamabad Capital Territory
*Indicative standard error indicates the precision of the estimate around the mean

Key Findings

Students may have some basic understanding of Geometry or Number Operations in Grade 4 Math, but Statistics and Probability items had lower scores than other subject areas. With the exception of Statistics and Probability, students' slightly lower performance is also observed in Algebra as compared to other Grade 4 Mathematics domains.

Punjab consistently outperform its provincial counterparts in all content domains of math. On the other hand, interestingly GB and ICT have the lowest math scores.

DISCUSSION

The study's aim was to appraise the National Achievement Test (NAT) and determine how effective it is as an evaluation tool. Using the Classical Test Theory (CTT) paradigm, the comprehensive psychometric research that was carried out for the years 2023 yielded significant findings about the test questions' discriminating power and difficulty indices. Several challenging items were found in both years, suggesting that the assessment needs to be revised or eliminated in order to provide a fair and accurate assessment of students' understanding.

Mathematical Item Difficulties

1 test item falls within the difficult index range's difficulty standards. This item was either relatively simple or slightly difficult and was deemed too challenging. As a result, this item should be changed or removed from the questionnaire. Understanding grouped data and how to interpret it in practical settings is a typical statistical topic that many students struggle with. Students often have difficulty with this since they have to relate abstract statistical ideas to real-world scenarios (Libman 2010).

Justification for Probability

This problem occurs when students find it difficult to comprehend complex probability concepts like probability distributions, conditional probability, or Bayes' theorem. For example, it could be difficult for students to understand the concept of independent events or how probability fundamentals are applied in complex scenarios. They may find it challenging to interact with and apply higher-level probability concepts in problem-solving scenarios due to a lack of basics or inadequate exposure to these concepts. The challenge occurs when students consider it difficult to formulate a methodical approach or road map for successfully solving probability questions. Without an organised approach, students might feel overwhelmed or disoriented during the problem-solving process, which might result in misunderstandings and ineffective methods (Mrope and Deogratias 2024).

Math Items Discrimination

Regarding discrimination, two additional items (domain: probability and real numbers), mainly in Grade 4 Math, were likewise deemed to be marginal and non-discriminant items and are unacceptable. Additionally, it has been found that item number M36 was an extremely difficult item question and is also a marginal discriminatory item. Real numbers are essential to computing and algorithms in many scientific and technical domains. Real numbers are essential in computer graphics for picture rendering, physical system simulation, and scaling, rotating and translating objects in a virtual environment. The finite representation of real numbers has drawbacks, including loss of precision and rounding errors. In order to prevent minor errors from becoming beyond control, algorithms must be designed to minimise these errors, guarantee stability, and converge to the right answer. In large-scale challenges, efficiency is also essential for managing real numbers, which lowers the amount of time and resources needed for calculation (Raikhola 2024).

Previous research has also shown that students' performance in mathematics, particularly in probability constructs, needs to be addressed. Regarding statistics, researchers demonstrate that students have difficulty identifying and utilising different measures of data concentration (correlational reasoning), proving variety (proof), and applying various concepts and principles of combination and permutation (combinatorial reasoning). Students struggle with these indicators because they are not familiar with handling proof-related concerns. Students hardly ever complete the mathematical task known as the question of proof (Puspitasari et al. 2019).

Interestingly, a few questions across numerous mathematical domains, including geometry and measurement, numbers and operations, algebra, and information handling were also challenging for students to comprehend. There is a lot of research that specifically addresses the factors that lead to students performing poorly in scientific classes. Mathematics performance is influenced by a variety of factors, and poor performance is the consequence of multifaceted collaboration among students, teachers, and educational factors (Koma 2025).

Furthermore, mastering formulas and techniques is solely insufficient for mathematics, a thorough understanding of the underlying ideas is required (Tomul et al. 2021). In this regard, students who lack an adequate basis in fundamental arithmetic, algebra, and geometry may find it challenging to comprehend disciplines like trigonometry or geometry. Lacking a solid grasp of the fundamentals, students might find it difficult to learn more complex subjects. Teachers serve as educational experts, ensuring students' academic achievement through efficient teaching-learning interactions in classrooms (Gnawali 2024). In this case, low performance results from the aforementioned cause of a lack of foundation. Additionally, the study showed geographical variations in item difficulty between students from ICT and AJK provinces, suggesting potential discrepancies in teaching strategies and curriculum implementation. Durry's (1999) research demonstrates an apparent relationship between curriculum and academic achievement, and it is widely acknowledged that children perform better when educational goals and objectives are well-defined. Other researchers also suggested that one has to reframe the curriculum's learning objectives, review preparing lessons in the context of the objectives, and reexamine appropriate teaching strategies and inquiry-based learning activities. At the classroom level, one ought to implement new teaching and evaluation strategies. In order to make it easier to choose the test items for the exit national examination, one should first compile the items bank with the specifics of each item's characteristics.

Cognitive Domain Wise

Grade 4 students' understanding as well as application of the mathematical concepts is assessed in the test and is apparent in this score. With an overall score, it seems that students are currently falling below the required level of competence in mathematics. If the curriculum prioritises rote memorising over problem-solving and reasoning, students might not acquire the skills required to succeed in the "applying" and "reasoning" domains (Hattie 2008). Moreover, teachers concentrate on the knowledge domain because many standardised tests emphasise recall and comprehension over application and reasoning (Popham 2010).

Justification

The scientific method encourages an application-focused educational atmosphere by emphasising the use of knowledge to formulate hypotheses, conduct experiments, and evaluate findings (Council et al. 2012). Many science tests emphasise how concepts are implemented through practical applications instead of only evaluating reasoning abilities (Wiggins and McTighe 2005). Voogt and Roblin (2010) framework emphasises the importance of application and problem-solving abilities in STEM education.

Comparisons of the Standard Error Percentage of Students' Math Marks Across Different Provinces (Cognitive Domain Wise) Justifications

Punjab got the highest grades and GB (Gilgit-Baltistan) obtained the lowest. This could be because of several factors, including disparities in the quality of education, socioeconomic status, government policies, and educational infrastructure (Awan 2018). Multiple factors such as student and teacher characteristics, home-based factors affect performance. Students should be able to interpret instructions and questions in addition to working independently and, most importantly, writing accurate answers. Being proficient in these abilities is essential for performing well in an exam. A child's academic achievement has a positive relationship with parental involvement in their early schooling (Hill and Craft 2003). Moreover, teachers' professional development procedures also have no impact on the practicalities of the classroom and school. PD implementation requirements cannot be adequately addressed by science teacher professional development methods. When the practical work was given less attention, the majority of science instructors lost motivation and became passive participants in professional development programs (Qureshi and Demir 2019).

Contents

Justification Math

Students may have some basic understanding of Geometry or Number Operations in Grade 4 Math, but Algebra, Measurement and Statistics and Probability items had lower scores than other subject areas. Regarding algebra, it is believed to be one of the main subjects that demands logical thinking from students. This encourages their skills and enables them to concentrate on arithmetic operations while emphasising the use of symbols to represent equations and to establish the relationships in mathematical operations (Ying et al. 2020). Moreover, the cognitive demands of the various algebraic operations and representations constitute mental obstacles. First-year algebra pupils lack the motivation to acquire the abilities essential to solve algebra word problems (Sevinç et al. 2024). This is partly because students try to avoid representational and transformational activities, generalising and arguing activities, and other activities (Muchoko et al. 2019).

Apart from Statistics and Probability, students' slightly lower performance is also observed in Algebra as compared to other Grade 4 Mathematics domains. Regarding the probability and the statistics domain, Shaughnessy (2007) explains the difficulties that students face when learning probability, such as misunderstandings and the abstract nature of probabilistic concepts. Numerous researchers have also emphasised how crucial it is for educators to possess a thorough understanding of statistics and probability in addition to being aware of any potential misunderstandings that students might have. Moreover, Burgess (2006) claimed that there are essential distinctions between the knowledge and comprehension required of teachers when teaching mathematics and statistics. Furthermore, the activities and representations teachers employ are crucial to students' successful learning of probability. The author further suggested various models and representations for teaching probability (such as Venn diagrams, tables and tree diagrams) (Kvatinsky and Even 2002). Educators must understand when each model or representation is appropriate, how they relate to one another, and a variety of examples and contexts that can demonstrate probabilistic concepts, properties, and theorems that pertain to probability.

Both professional experiences and personal knowledge have an impact on science instructors' professional development and knowledge construction (Begum et al. 2020). Qureshi and Demir (2019) suggested that the quality of the curriculum and textbooks need to be examined in Pakistan. It was highlighted that the shortage of personnel and inadequate recruitment procedures meant that the textbook authors lacked the necessary subject-matter competence, which in turn affected the quality of the texts. These authors lack the pedagogical knowledge to link curriculum to students' learning outcomes and are not prepared to write textbooks.

CONCLUSION

This study highlights the pressing need to reevaluate the national achievement test, particularly in terms of its development and application. It also emphasises how crucial it is to review the national achievement test, especially with reference to its development and administration. Although a majority of the questions in the mathematics tests seem appropriate and relevant, the examination of the difficulty and discrimination indices shows that there are some items in mathematics (particularly in the areas of geometry, measurement, and algebra) that should be revised or eliminated because of their difficulty or lack of discriminatory characteristics. This underlines how imperative it is to ensure extensive domain coverage, take regional variances into consideration, and improve item construction in order to improve the standard and precision of the assessments.

RECOMMENDATIONS

Variability in item difficulty and discriminating indices between regions were identified during the development phase study. It is therefore suggested to consider these variations into account when constructing test items and to use

Automated Test Assembly (ATA) to choose and arrange items in accordance with predetermined psychometric requirements. To eliminate or modify the problematic items, the NAT standard configuration must be created based on the post-exam analysis.

ACKNOWLEDGMENTS

The authors would like to thank the National Assessment Wing, the Pakistan Institute of Education (PIE), the Ministry of Federal Education and Professional Training, and the National University of Modern Languages (NUML), Islamabad, Pakistan.

REFERENCES

Awan N 2018. Comparative study of availability and quality of physical facilities in public and private schools in the Punjab. *Journal of Elementary Education,* 28(1): 99-107.

Begum Z, Rahman A, Majoka MI 2020. Contribution of professional efficacy and demographic factors towards performance of secondary schools teachers. *Dialogue (Pakistan),* 15(1): 137-146.

Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL 2018. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health,* 6: 149.

Burgess T 2006. A Framework For Examining Teacher Knowledge As Used In Action While Teaching Statistics. *Proceedings of the Seventh International Conference on Teaching Statistics,* International Statistical Institute. Salvador, Brazil, 2-7 July.

Drury DW 1999. *Reinventing School-Based Management: A School Board Guide to School-Based Improvement*. Alexandria, VA: National School Boards Association.

D'Sa JL, Visbal-Dionaldo ML 2017. Analysis of multiple choice questions: Item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education,* 9(3): 109-114.

Gnawali YP 2024. Causes of poor performance in Mathematics at School Education Examination (SEE). *Ganeshman Darpan,* 9(1): 71-78.

Hattie J 2008. *Visible Learning: A Synthesis Of Over 800 Meta-Analyses Relating To Achievement*. Routledge.

Hill NE, Craft SA 2003. Parent-school involvement and school performance: Mediated pathways among socioeconomically comparable African American and Euro-American families. *Journal of Educational Psychology,* 95(1): 74.

Hopfenbeck TN 2016. The power of PISA: Limitations and possibilities for educational research. *Assessment in Education: Principles, Policy & Practice,* 23(4), 423–426. https://doi.org/10.1080/0969594X.2016.1188374

Kohler C, Robitzsch A, Hartig J 2020. A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics,* 45(3): 251-273.

Kolawole EB 2005. *Measurement And Assessment In Education.* Lagos: Bolabay Publication.

Kolomuc A, Demir Saglam M 2025. Development of Science Attitude Scale for middle school students with special needs studying in Inclusive Education Program. *International Journal of Science and Mathematics Education*, 1-24.

Koma TG 2025. An Investigation Of Grade 12 Mathematics Teachers' Competencies In The Topic Of Probability. From <https://www.scribd.com/document/938977982/Training-Manual-Probability> (Retrieved on 25 December 2025).

Kvatinsky T, Even R 2002. Framework For Teacher Knowledge And Understanding About Probability. *Proceedings of the Sixth International Conference on Teaching Statistics (CD).* International Statistical Institute, Cape Town, South Africa, 7-12 July 2002.

Libman Z 2010. Integrating real-life data analysis in teaching descriptive statistics: A constructivist approach. *Journal of Statistics Education,* 18(1): 187. https://doi.org/10.1080/10691898.2010.11889495

Mahjabeen W, Alam S, Hassan U, Zafar T, Butt R, Konain S, Rizvi M 2017. Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University,* 13(4): 310-315.

Meyer HD, Benavot A 2013. PISA, Power, And Policy: The Emergence Of Global Educational Governance (Illustrated Edition). Oxford, United Kingdom: Symposium Books Ltd. From <https://books.google.com.pk/books?id=fqFwCQAAQBAJ > (Retrieved on 7 December 2025).

Morris W, Holmes L, Choi JS, Crossley S 2025. Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education,* 35(2): 559-586.

Mrope FM, Deogratias E 2024. Analysis of students' difficulties in solving probability questions: The case of advanced certificate of secondary education examination in Tanzania. *Union: Jurnal Ilmiah Pendidikan Matematika,* 12(3): 494-506.

Mukherjee P, Lahiri SK 2015. Analysis of multiple choice questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. *IOSR J Dent Med Sci*, 14(12): 47-52.

Muchoko C, Jupri A, Prabawanto S 2019. Algebraic visualization difficulties of students in junior high school. *Journal of Physics,* 1157(3): 32-108.

Musa A, Shaheen S, Elmardi A, Ahmed A 2018. Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University. *Khartoum Medical Journal,* 11(2): 105-112.

Nurtamam ME, Santosa TA, Ilwandri I, Rahman A 2023. Meta-analysis: The effectiveness of Iot-based flipped learning to improve students' problem solving abilities. *Edumaspul: Jurnal Pendidikan,* 7(1): 1491-1501.

Pakistan Institute of Education 2024. National Achievement Test. From <https://pie.gov.pk/SiteImage/Downloads/NAT%202023%20Findings%20Report%2006.03.2024%20-Final-%20v6%20.pdf> (Retrieved on 7 December 2025).

Popham WJ 2010. *Everything School Leaders Need To Know About Assessment.* Thousand Oaks, CA, USA: Corwin Press.

Puspitasari N, Afriansyah E, Nuraeni R, Madio S, Margana A 2019. What are the difficulties in statistics and probability? *Journal of Physics: Conference Series,* 1234, 012034. https://doi.org/10.1088/1742-6596/1234/1/012034

Qureshi AM, Demir K 2019. A comparative review of the literature on Pakistani Science teachers' professional development. *Science Education International,* 30(3): 222-235.

Quinn Helen, Schweingruber Heidi, Keller Thomas 2011. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC, USA: National Academies Press.

Rahman NA, Rosli R, Rambely AS, Halim L 2021. Mathematics teachers' practices of STEM Education: A systematic literature review. *European Journal of Educational Research,* 10(3): 1541-1559.

Raikhola SS 2024. Mathematical foundations of real numbers and its application in computation. *Swarnadwar,* 4(1): 71-86.

Sahoo DP, Singh R 2017. Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. *International Journal of Research in Medical Sciences,* 5(12): 5351-5355.

Sevinç S, Bostan MI, Çakiroglu E 2024. First-grade students' strategy choices in addition word problems. *Education and Science,* 49(220): 59-81.

Shaughnessy JM 2007. Research on statistics' reasoning and learning. In: FK Lester Jr. (Ed.): *Second Handbook Of Research On Mathematics Teaching And Learning.* Charlotte, NC, USA: Information Age Publishing, pp.957-1009.

Sim S-M, Rasiah RI 2006. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals-Academy of Medicine Singapore,* 35(2): 67.

Tomul E, Önder E, Taslidere E 2021. The relative effect of student, family and school-related factors on math achievement by location of the school. *Large-scale Assessments in Education,* 9(1): 22.

Voogt J, Roblin NP 2010. 21st century skills. *Discussienota. Zoetermeer: The Netherlands: Kennisnet,* 23(3): 2000.

Wiggins GP, McTighe J 2005. *Understanding By Design*. Alexandria, VA, USA: Ascd.

Yang Y, Maeda Y, Gentry M 2024. The relationship between mathematics self-efficacy and mathematics achievement: Multilevel analysis with NAEP 2019. *Large-scale Assessments in Education,* 12(1): 16.

Ying CL, Osman S, Kurniati D, Masykuri ES, Kumar JA, Hanri C 2020. Difficulties that students face when learning algebraic problem-solving. *Universal Journal of Educational Research,* 8(11): 5405-5413.

Paper received for publication in November, 2025
Paper accepted for publication in February, 2026