International Journal of Assessment Tools in Education
2026, Vol. 13, No. 1, 108–122
https://doi.org/10.21449/ijate.1570567
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Expert judgment vs. psychometric analysis: Enhancing high school exam development in Albania

Eralda Gjika[1*], Lule Basha[2], Afërdita Alizoti[3], João Paulo Araujo Lessa[4]

[1]University Metropolitan Tirana, Faculty of Computer Engineering and IT, Tirana, Albania
[2]University of Tirana, Faculty of Natural Science, Tirana, Albania
[3]Centre of Educational Services, Tirana, Albania
[4]University Tuiuti do Paraná, Brazil

ARTICLE HISTORY
Received: Oct. 21, 2024
Accepted: Sep. 23, 2025

Keywords:
E-assessment,
Item analysis,
CTT,
Expert judgment,
High school exams.

Abstract: In Albania, high school graduates undergo State Matura exams that determine university entrance through a merit-based system. Since 2006, these pencil and paper exams have included three mandatory exams: Albanian language and literature, Mathematics, and a foreign language, plus one elective exam from a list of eight. Each exam totals 60 points: 20 points for multiple-choice items and 40 points for open-response items (structured and essay-type) covering Albanian literacy and a foreign language. Historically, item difficulty has been set primarily by expert judgment, with limited psychometric validation, and student proficiency has been computed via Classical Test Theory (CTT) on a 4-10 decimal scale. This study highlights the lack of psychometric analysis in Matura exams, emphasizing the need for improved assessment methods. By focusing on the limitations of relying solely on expert judgment in an era of technological innovations, we address the challenges posed by insufficient historical item parameters. To support expert judgment, we present a ShinyApp that integrates exam data to provide fast, transparent, and replicable psychometric insights. The tool demonstrates how technology can support evidence-based decision-making and contribute to modernizing Albania's e-assessment framework.

INTRODUCTION

Assessments play a crucial role in guiding students toward their college or university paths, preparing them with the competencies and expertise required to succeed in their careers. On the other hand, they act as significant indicators of academic progress. These data are valuable for educators and policymakers who can benefit by identifying areas of strength and weakness in the curriculum, teaching, and learning methodologies. Timely and appropriate actions can greatly contribute to the overall success of the educational program (OECD, 2024a). Among state assessments, high school examinations play an important role in the educational journey of students, often determining their eligibility for higher education. In many educational systems as in Albania, these exams serve as an important bridge to college or university admissions by shaping this way students' career trajectories. In this study, we analyze one of

*Corresponding author at: Eralda GJIKA eraldagjika@gmail.com University Metropolitan Tirana, Faculty of Computer Engineering and IT, Tirana, Albania

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/
e-ISSN: 2148-7456

the elective subject exams from the Albanian State Matura, a national high school graduation assessment. The examination process is centrally regulated by the Center for Educational Services (CES), which is responsible for item creation, exam design, administration, marking, scoring, and the dissemination of final results to students.

State Matura examinations (previously known as high school examinations) were introduced in 2006 as a pre-university examination system. Apart from the evaluation of student performance at the end of high school, these results are important for determining students' eligibility for various university degree programs in Albania. Ever since their initial presentation in 2006, these exams have been administered in a traditional pencil and paper format. The Matura exams are organized into two main categories: mandatory and elective subjects. The mandatory exams include Mathematics, Albanian Language and Literature, and a Foreign Language (one among the list: English, French, German, Italian, Spanish, and Turkish). In addition, students must choose one subject (elective) from a list of eight subjects (Biology, Chemistry, Economics, Philosophy and Sociology, Physics, Geography, History, Citizenship, and Psychology). This elective subject exam score will influence their university entrance. There have been significant changes in the system since 2006, evolving continuously during its development. In 2006 and 2007, elective exams were all MCQs constructed with a total of 50 points. From 2006 to 2017, mandatory subject exams comprised 25 questions (50 scores in total), with 13 questions featuring multiple-choice questions (MCQ) with a maximum score of one and 12 questions requiring short responses (SAQ-Short Answer Questions) with a total of 37 scores. From 2008 to 2018, for mandatory exams, the format shifted to a combination of MCQ (10 questions with a total allocation of 10 scores) and SAQ (20 questions with a total allocation of 30 scores). Moreover, the system underwent further important changes during the COVID-19 pandemic. Starting from the State Matura in 2019, all tests (mandatory and elective) carry a maximum of 60 scores, with 20 scores designated for MCQ and 40 scores for SAQ, including structured and essay-type questions in Albanian literacy and foreign language. According to exam development guidelines by CES, item difficulty combination in a test (based on score) should follow a ratio of 40:40:20, corresponding to low, medium, and high difficulty levels, respectively (MoES, 2019). The exam duration is 2 hours and 30 minutes for all exams (mandatory/elective). Each exam is administered on a different day over a period of one month. The mandatory exams are conducted first, followed by the elective exams. Students receive their exam results within one week after each exam is administered. Once results are released, students begin applying for their university studies. Between 2006 and 2019, the passing cut score for the Matura exams in Albania was determined annually based on student performance in that year's examination. This adaptive approach aimed to ensure a fair and context-sensitive evaluation of student achievement. As of 2020, students must attain 25 percent of the test scores to achieve passing status (MoES, 2023). Student proficiency levels are assessed using a decimal grade scale ranging from 4 to 10, which is calculated based on the 25% cut score (equivalent to a 15 score out of the total score of 60). The exam development process has relied heavily on subjective judgments by experienced educators. To date, there has been no formal reporting on exam performance and student achievement outcomes. Establishing a more rigorous evaluation framework would provide stronger evidence of the reliability and validity of exam items and offer clearer insights into their impact on students' progression to university (Maghnouj *et al*., 2020).

Since the inception of the State Matura exams in Albania, there have been efforts to establish a framework for assessing the reliability and accuracy of high school exams. However, due to the lack of a structured system, several intermediate steps have not had sufficient historical documentation. Question bank creation has been in focus since the early stages of the State Matura; however, the absence of statistical data and psychometric analyses of the tests and questions highlights the importance of such evidence to ensure that these banks can be employed to their full potential (Wright & Bell, 1984; Bjorner *et al*., 2007).

More specifically, there has not yet been a systematic record of the psychometric performance of multiple-choice questions (MCQs) and written questions (SAQs) over the years. This limitation restricts item developers and subject experts from effectively rekeying or reconstructing new questions based on previously used ones. The lack of this historical context makes it challenging to establish a clear relationship between new questions and those previously employed in tests, while also impeding the ability to pilot questions and anticipate outcomes. On the other hand, item banks are key in developing e-assessments, as they provide a structured repository of validated questions and allow for piloting and refining items to ensure accurate and reliable assessment outcomes. Developing e-assessments automates the process by improving the accuracy of results and increasing efficiency in terms of time management (Ke *et al*, 2010; Schwarz & Gjika, 2023; Schwarz *et al*, 2023).

As the study progresses, we aim to investigate several key questions:

To what extent do the item development team's evaluations support the conclusion that the exams are well-constructed and meet accepted standards of reliability? Furthermore, is there a need to reassess the current guidelines and frameworks to improve the exam development process and ensure stronger alignment with the national curriculum?

In this phase of the study, we undertake a case study of one elective subject exam from the Albanian State Matura. We conduct a thorough statistical and psychometric analysis, as the data allows, addressing the following questions:

- To what extent does the observed item difficulty distribution conform to the intended 40:40:20 ratio of low, medium, and high difficulty levels following exam administration?
- How do student performance outcomes differ between multiple-choice questions (MCQs) and short-answer questions (SAQs)?
- What underlying patterns can be identified through item clustering techniques, and how might these inform future exam design?

This study's findings will provide valuable insights into the expert judgment and psychometric evaluation, specifically within the framework of Classical Test Theory (CTT).

The paper is organized as follows: In Section 2, we review similar frameworks in the literature, as well as comparisons of international and regional experiences. Section 3 describes the exam format and procedures. In Section 4, we present the statistical and psychometric analysis of the exam under consideration. Section 5 discusses the findings and potential improvements to the framework and concludes the article.

LITERATURE REVIEW

Reform Impacts

Since joining the Programme for International Student Assessment (PISA) in 2001, Albania has been actively engaged in improving its educational standards and outcomes (Shahini, 2021). The participation of Albania in PISA has prompted significant reforms aimed at aligning its education system with international benchmarks. In an effort to enhance student performance, Albania has been adapting its assessment formats to mirror the structure of the PISA tests, which emphasize critical thinking, problem-solving, and the application of knowledge in real-world contexts. This alignment not only aims to improve students' competencies in key subjects like reading, mathematics, and science but also to better prepare them for the challenges of a rapidly changing global landscape (OECD, 2024b).

Similar reform efforts have been undertaken across the region to align more closely with the PISA framework, recognizing the importance of standardized assessments in measuring student performance and fostering educational improvement. In their study, Firescu *et al* (2022) examine recent changes to the Romanian education system, particularly the reduction of the curriculum and restructuring of the National Evaluation exam to resemble PISA tests. The findings indicate progress in high school efficacy due to grade inflation, but highlight that most high schools perform poorly, emphasizing the need for more robust evaluation metrics to assess educational institutions' effectiveness.

Similar findings have been reported by other authors in their studies. Kreitchmann *et al*. (2019) explore discrepancies between expert-rated item difficulty and empirical difficulty indices, highlighting the importance of integrating psychometric validation into test development. Likewise, Suto and Greatore (2008) examine the reliability of expert judgments in high-stakes exam settings, emphasizing the need for statistical calibration to ensure fairness and consistency. Wonde *et al*. (2024) found a moderate correlation (R = 0.57) between expert-rated and actual item difficulty in a medical licensure exam, highlighting the need to complement expert judgment with psychometric validation. These studies underscore the relevance of combining expert input with rigorous psychometric analysis, an approach central to our investigation of the Albanian Matura exams.

Psychometric Analysis Importance

In our study, we recognize the importance of integrating historical perspectives on educational assessment with contemporary practices to enhance validity and reliability.

Brownstein *et al*. (2019), in their study, emphasize the role of expert judgment in statistics, using case studies to illustrate its importance. The authors call for greater involvement of statisticians in research to ensure that judgments are informed by objective data and uncertainty assessments. Combining psychometric expertise with expert judgment is essential for developing reliable and accurate exams. Reliance solely on expert judgment can sometimes falter due to rapid technological evolution and the emergence of new sources of information that experts may not fully account for. To align with the rapid technological developments and trends in e-assessment, it is crucial that automated exam marking and scoring be performed swiftly and efficiently, enabling timely feedback for students. In the following section, we will analyze in detail how these two critical factors interact in the current landscape of educational assessment in Albania.

As highlighted in recent discussions on the evolution of assessment (Brown, 2022), there has been a significant reliance on expert judgment and the development of statistical methods such as classical test theory and item response theory that have improved test validity and reliability. This review traces the evolution of educational assessment from historical practices to contemporary methods, highlighting the shift from traditional paper-and-pencil tests to computer-based assessments that utilize adaptive scoring. In their latest studies (Schwarz *et al*., 2023; Schwarz & Gjika, 2023) introduce a modernized and user-friendly interface for contemporary psychometric analysis practices. They provide insights into the workflow design and functionality of the data pipeline, proposing an innovative approach to psychometric analysis in the realm of digital assessments. Their demonstration emphasizes the pipeline's efficiency and reliability, showcasing its ability to generate psychometric results in a matter of minutes by adopting an open-source model that shows transparency and accessibility in psychometric analysis.

CASE STUDY

For this study, we analyzed data from one administration of the Matura examination that aligns with the format used between 2019 and 2023. The selected dataset reflects a representative elective subject and was chosen based on relevance to the study’s objectives.

Exam Format

Based on the indicative program, the fulfillment of the subject competencies and key competencies that a student should master throughout the development of the subject is achieved through 2 main topics. In this study, ‘Topic 1’ and ‘Topic 2’ refer to the two main domains of the subject, each of which is structured into several sub-domains. The indicative program of the State Matura is implemented correctly by specialists and teachers involved in drafting the subject exam questions. Specialists involved in drafting exam questions for this subject are expected to align the content strictly with the topics outlined in the indicative program of the State Matura. This study does not aim to critique or evaluate the State Matura itself, but rather to illustrate, using a sufficiently representative dataset, how expert judgment and psychometric analysis can complement one another in pursuit of optimal assessment quality.

Test Structure

The test for an elective subject has a total of 60 scores, which are divided as follows:

- 20 multiple-choice questions (MCQ), which have a total of 20 scores (33.33% of scores)
- 14 short answer questions in the form of open-ended questions (SAQ), which have a total of 40 scores (66.67 % of scores)
- The elective subject test is structured in such a way that all three levels of difficulty are included:
- Basic (low) level (40% of the total score, 24 score overall),
- Intermediate (medium) level (40% of the total score, 24 score overall), and
- Advanced (high) level (20% of the total score, 12 score overall).

At the same time, the distribution of scores in the test is also dependent on the percentages occupied by each level of difficulty. Below is the Table 1 of test specifications, which shows the distribution of scores based on the percentage of topics in the indicative program and levels of difficulty.

Table 1. *Item Difficulty distribution among Topics.*

			Difficulty Level		
Primary Topics	Weight (%)	Score	Low (40%)	Medium (40%)	High (20%)
Topic 1	40	24	10	10	4
Topic 2	60	36	14	14	8
Total	100	60	24	24	12

Topic 1 represents 40% of the weight of the program and is represented in the test with approximately 24 scores, which are distributed by judgmental approach in the ratio 40:40:20 of item difficulty, while Topic 2 represents 60% of the program, has 36 scores with the same approach of distribution of item difficulty.

Types of Questions

The exam includes both objective and constructed-response questions that assess students' knowledge and skills across almost all topics. Constructed-response questions, such as structured short-answer items (SAQs), require students to analyze visual data (such as maps, figures, or graphs) and provide concise, structured answers. While these questions are not opinion questions, they are interpretative and synthesizing questions that demand information given to be worked through, distinguishing them from purely objective forms.

METHODOLOGY

The dataset includes information from all students (N = 2018) who sat for the selected elective exam during a single administration year. Regarding the exam structure: the test consists of 20 multiple-choice questions (MCQ) and 14 short-answer questions (SAQ), totaling 34 questions administered over 2.5 hours. The minimum average score for MCQ questions was 0.19, and the maximum average score was 0.81, with a standard deviation range from 0.39 to 0.5. The SAQ questions had a combination of 2, 3, or 4 subparts. Specifically, 3 questions had 2 sub-parts, 9 questions had 3 sub-parts, and 1 question had 4 sub-parts. The minimum average score for SAQ items was 0.23, and the maximum average score was 1.39, with a standard deviation range from 0.6 to 1.28.

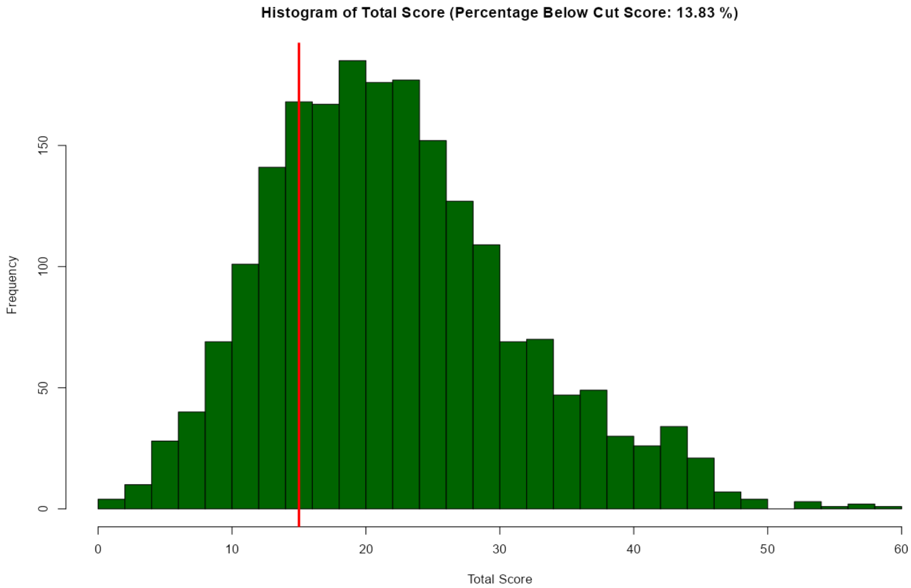

Figure 1. *Score distribution Histogram, and Cut-Score.*

The exam score distribution exhibits a slight right-skewed distribution from the graph below, indicating a considerable proportion of students who achieved low scores on the exam. However, the cut score (which was 25% of the scores) showed that 86.17% of students passed the exam. Most of the statistical and psychometric output for the exam analysis was obtained as download objects from the TIA Shiny application developed by Gjika *et al*. (2025).

Reliability Analysis of Exam Results

Two of the most commonly used reliability measures in educational and psychological research are: the coefficient Cronbach's α and the original Greatest Lower Bound (GLB), which are based on classical test theory.

Cronbach's α coefficient may be easily calculated in a matrix of scores using the formula in Equation 1 (Cronbach, 1951).

$$A=\left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sigma_{y_i}^2}{\sigma_y^2}\right) \quad (i,j=1,\dots,n) \tag{1}$$

where:

n represents the number of items in the test

$\sigma_{y_i}^2$ the variance associated with each item i

σ_y^2 the variance associated with the total scores

While the GLB is defined as in Equation 2:

$$GLB=1-\frac{TR(Ce)}{Var(X)} \tag{2}$$

, where TR(Ce) is the trace of the error inter-item covariance matrix and Var(X) is the variance of the sum score. Both coefficients may be easily calculated for a large sample size using the "psych" package in R (Revelle, 2024), which was incorporated in our Shiny app.

While Cronbach's Alpha offers insights into internal consistency, Greatest Lower Bound (GLB) was included as a measure of reliability because it offers a conservative estimate, particularly suited for mixed-format assessments that may be multidimensional or congeneric (Ten Berge

& Sočan, 2004; Sijtsma, 2009). Given that the selection exam comprises 60% multiple-choice and 40% short-answer items. The findings applied to our data indicated a Cronbach's Alpha of 0.84, which suggests a high level of internal consistency among the exam items. This level of reliability is generally considered acceptable, indicating that the questions within the exam are measuring a cohesive construct. Additionally, the GLB score was found to be 0.91, reflecting an even stronger reliability estimate. This result implies that the exam possesses a robust measurement framework. From a psychometric point of view, this suggests that the items are aligned and that the subject item development team is experienced in item preparation, ensuring consistency in their assessments.

Item Difficulty

The item difficulty index is a key characteristic in classical test theory that should be determined and examined at the outset of item analysis. In our study, the primary focus will be on comparing the item difficulty estimated by item developers' experts' judgment with the empirical item difficulty observed after the items have been administered. This comparison aims to assess the accuracy of the judgmental approach against empirical data. Classical test theory considers item difficulty as the primary characteristic to be determined, reflecting the alignment between item difficulty and student abilities. However, a mismatch between item difficulty and student performance may indicate item-level issues, best assessed through the discrimination index rather than overall test reliability. In these cases, when items exhibit discrepancies between estimated and actual difficulty, further examination for potential deficiencies or biases may be required (Warburton & Conole, 2003; McAlpine & Hesketh, 2003).

In our data, items are classified into two categories: Multiple-Choice Questions (MCQs) and Short Answer Questions (SAQs). Each category contains a combination of items with varying levels of difficulty: low, medium, and high. The item development team, composed of high school teachers and CES representatives, prepares exams following strict guidelines, including the distribution of item difficulty and alignment with domains and sub-domains. Items are classified by experts when added to the bank, and in the final stage, the team selects from this bank to assemble the exam (MoES, 2019; MoES, 2023).

The overall distribution of item difficulty across the entire exam is designed to follow a specific rate: 40% score of the items are of low difficulty, 40% are of medium difficulty, and 20% are of high difficulty. However, it is important to note that this distribution rate is calculated based on the item scores, rather than the number of items. This means that the proportion of low, medium, and high difficulty items is determined by the contribution of each item's score to the total score of the exam, ensuring that the difficulty distribution reflects the overall scoring weight of the exam. Table 2 shows the item statistics for MCQ and SAQ items in the exam, as well as the expert judgment item difficulty classification before the administration.

In our data analysis, we consider items as whole units and not their subparts. Therefore, the analysis of item difficulty presented in Table 3 is based on the items themselves and not on the scores associated with each item. This approach ensures that our examination of item difficulty captures the full range of item characteristics, aligning with the methodology employed by the item development team during the exam creation process. We can clearly observe how the score percentage (40:40:20) between the three categories of item difficulty is distributed across the entire exam and the total number of items. The table also highlights the distribution between Multiple-Choice Questions (MCQs) and Short Answer Questions (SAQs), providing a clear comparison of difficulty levels for each item type.

Table 2. *Item statistics for the MCQ and SAQ.*

Item ID	mean	*sd*	median	min	max	range	Expert Judgment Difficulty
MCQ_1	0.709	0.454	1	0	1	1	LOW
MCQ_2	0.623	0.485	1	0	1	1	MEDIUM
MCQ_3	0.814	0.389	1	0	1	1	MEDIUM
MCQ_4	0.264	0.441	0	0	1	1	HIGH
MCQ_6	0.737	0.44	1	0	1	1	LOW
MCQ_8	0.189	0.392	0	0	1	1	MEDIUM
MCQ_9	0.198	0.398	0	0	1	1	HIGH
MCQ_11	0.375	0.484	0	0	1	1	LOW
MCQ_13	0.465	0.499	0	0	1	1	LOW
MCQ_15	0.684	0.465	1	0	1	1	LOW
MCQ_16	0.778	0.416	1	0	1	1	LOW
MCQ_17	0.492	0.5	0	0	1	1	LOW
MCQ_18	0.604	0.489	1	0	1	1	LOW
MCQ_19	0.769	0.422	1	0	1	1	LOW
MCQ_20	0.599	0.49	1	0	1	1	LOW
MCQ_21	0.558	0.497	1	0	1	1	LOW
MCQ_27	0.732	0.443	1	0	1	1	LOW
MCQ_28	0.446	0.497	0	0	1	1	LOW
MCQ_29	0.578	0.494	1	0	1	1	MEDIUM
MCQ_30	0.279	0.449	0	0	1	1	LOW
SAQ_5	0.749	0.751	1	0	2	2	LOW
SAQ_7	1.262	0.809	1	0	3	3	MEDIUM
SAQ_10	0.531	0.899	0	0	3	3	MEDIUM
SAQ_12	0.305	0.607	0	0	2	2	LOW
SAQ_14	0.348	0.704	0	0	3	3	HIGH
SAQ_22	0.815	0.941	1	0	3	3	MEDIUM
SAQ_23	1.391	1.28	1	0	4	4	HIGH
SAQ_24	1.165	1.053	1	0	3	3	MEDIUM
SAQ_25	1.358	0.749	2	0	2	2	MEDIUM
SAQ_26	1.379	1.165	1	0	3	3	LOW
SAQ_31	0.687	0.939	0	0	3	3	LOW
SAQ_32	0.409	0.791	0	0	3	3	MEDIUM
SAQ_33	1.048	1.117	1	0	3	3	MEDIUM
SAQ_34	0.233	0.625	0	0	3	3	HIGH

Table 3 illustrates how item difficulty distribution appears in the exam, both in terms of the number of items and the score distribution for each category. One thing to consider by observing Item ID in Table 2 is that the item ID identifies the type of the item (MCQ or SAQ) and also the order the item will appear in the exam (i.e., MCQ_1 or SAQ_12). So, the exam can be considered as a two-paper exam with one MCQ and another SAQ, but the papers are not clearly distinguished in the test.

Table 3. *Item Difficulty distribution based on the number of items and score.*

Item Difficulty		No. of Items	No. Items (%)	Scores	Score (%)
Low		18	52.94	24	40.00
	MCQ	14	41.18	14	23.33
	SAQ	4	11.76	10	16.67
Medium		11	32.35	24	40.00
	MCQ	4	11.76	4	6.67
	SAQ	7	20.59	20	33.33
High		5	14.71	12	20.00
	MCQ	2	5.88	2	3.33
	SAQ	3	8.82	10	16.67
Total		34	100	60	100

Table 4 summarizes the results of item difficulty prior to and post exam administration. It may be used in support of the item difficulty performance before (expert judgment classification) and after (empirical data). Item difficulties for MCQs and SAQs were calculated as the proportion of examinees answering correctly (p-value), using functions from the psych, CTT, and ltm R packages (Revelle, 2024).

Table 4. *Item Difficulty Categories (Prior vs Post (%) analysis by number of items).*

	Prior (%)			Post (%)		
Item Type	Low	Medium	High	Low	Medium	High
MCQ	70	20	10	30	50	20
SAQ	29	50	21	0	50	50
Total	53	32	15	18	50	32

Table 4 (Prior column) item difficulty distribution shows how MCQ and SAQ items are distributed following the number of items' difficulty based on expert judgment. It shows a distribution overall of 53%-32%-15% among corresponding levels of difficulty, Low-Medium-High. Table 4 (Post column) reports how this distribution shows after the exam is administered. We clearly observe a shift in the percentage from low to Medium and High levels, 18%-50%-32%.

We observe that the ratio distribution remains almost the same, approximately 50(L)-30(M)-20(H), but the classes have now shifted to 50(M)-30(H)-20(L). To quantify the alignment between these two sources, we applied a metric we refer to as Item Difficulty Consistency (IDC), conceptually adapted from the structure of a confusion matrix commonly used in classification tasks (Saito & Rehmsmeier, 2015). By Item difficulty consistency, we refer to the stability regarding how difficult items in a test are perceived to be, particularly before and after statistical estimation or modeling. For this analysis, we compared the initial estimates, based on expert judgment, with estimates obtained through statistical item analysis.

$$Item\ Difficulty\ Consistency\ (IDC) = \frac{N_L + N_M + N_H}{N} \qquad (3)$$

In Equation 3:

- N_L is the number of items correctly classified at level *Low*
- N_M is the number of items correctly classified at level *Medium*
- N_H is the number of items correctly classified at level *High*
- N is the total number of items in the test

The IDC values will range between 0 and 1, where:

- 1 indicates perfect consistency (all items retained their original classifications).
- Values near 0 reflect low consistency (many items were misclassified).

- Values near 0.5 suggest moderate consistency (indicating partial alignment between expert judgment and statistical outcomes)

A high degree of consistency will indicate that the items are consistently perceived as easy, moderate, or difficult, regardless of the estimation method used. In our data analysis, we observe that the calculated consistency is close to 0.42. This value is below the expected average (0.5) and suggests a moderate level of agreement between the expert judgment and statistical analysis. This value highlights the need for further investigations on expert judgment use and analyzing specific items and their response patterns. Table 5 shows the classification of items for the full exam (MCQ and SAQ), Prior and Post exam administration.

Table 5. *Prior and post item difficulty classification.*

Difficulty Level		Post		
		Low	Medium	High
Prior	Low	5	10	3
	Medium	1	6	4
	High	0	1	4

In association with it, Figure 2 shows the heatmap of the percentage of items (MCQ) that were classified in a different class (Low, Medium, High) after the exam administration. By observing the heatmap for MCQ items, we notice a considerable percentage (29.4%) of items that moved from Low difficulty to Medium after the exam was administered. Almost 41.2% changed classification by moving one category up (from Low-Medium, Medium to High). 8.8% changed their classification range by moving two categories up (Low to High). The percentage of items moving from a high range to a lower one is approximately 5.8% (High- Medium and Medium-Low). No items were observed to fall from the High range to the Low range. For these changes in the classification category, further investigations are needed. The high percentage of items shifting from low to high difficulty suggests a need for further evaluation of item quality and alignment with learning objectives. However, this shift may also reflect contextual factors such as the specific test selected for analysis or the limited scope of data, which was drawn from a single school. These considerations highlight the importance of interpreting item performance within the broader assessment context.

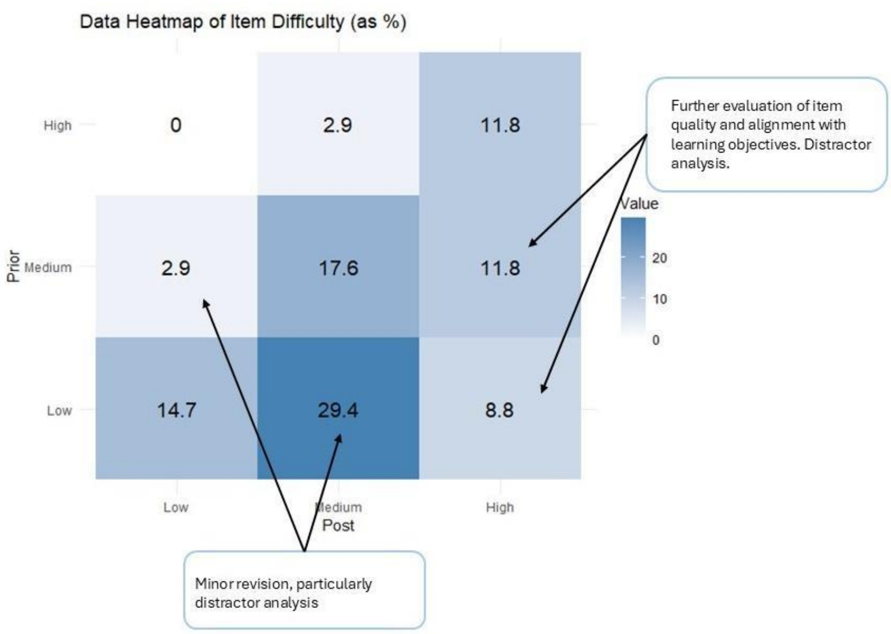

Figure 2. *Heatmap of Item Difficulty classification change (% of items) for MCQ item type before and post exam administration.*

Item Discrimination

Item discrimination is a measure used to show how items differentiate between high-ability and low-ability students and plays a crucial role in evaluating test quality. In this study, for dichotomously scored items (MCQs scored correct/incorrect), item discrimination was calculated using the point-biserial correlation between item score and total test score. For polytomously scored SAQs, item discrimination was calculated using the Pearson correlation between the item score and the total test score (Ebel & Frisbie, 1991).

Ebel's criteria and guidelines for categorizing discrimination indices are based on the correlation between item scores and total test scores (Ebel, 1972). Strong Discrimination: Items with discrimination indices (also known as item-total correlations) above 0.40 are considered to have strong discrimination. These items effectively differentiate between high-ability and low-ability individuals and are desirable for most assessments.

Moderate Discrimination: Items with discrimination indices between 0.20 and 0.40 are categorized as having moderate discrimination. While these items still contribute to differentiating between ability levels, they may not be as effective as strongly discriminating items and may require further evaluation or revision.

Weak Discrimination: Items with discrimination indices below 0.20 are classified as having weak discrimination. These items may not effectively differentiate between high and low ability, and their contribution to the overall test's discriminative power is limited. Generally, these items may need to be removed or improved.

Figure 3 shows the behavior of MCQ and SAQ item difficulty and discrimination index overall in the pool of items used in the exam. We observe that SAQs tend to have higher discrimination levels compared to MCQs, which is expected after the shift of the item difficulty in Post post-administration range.

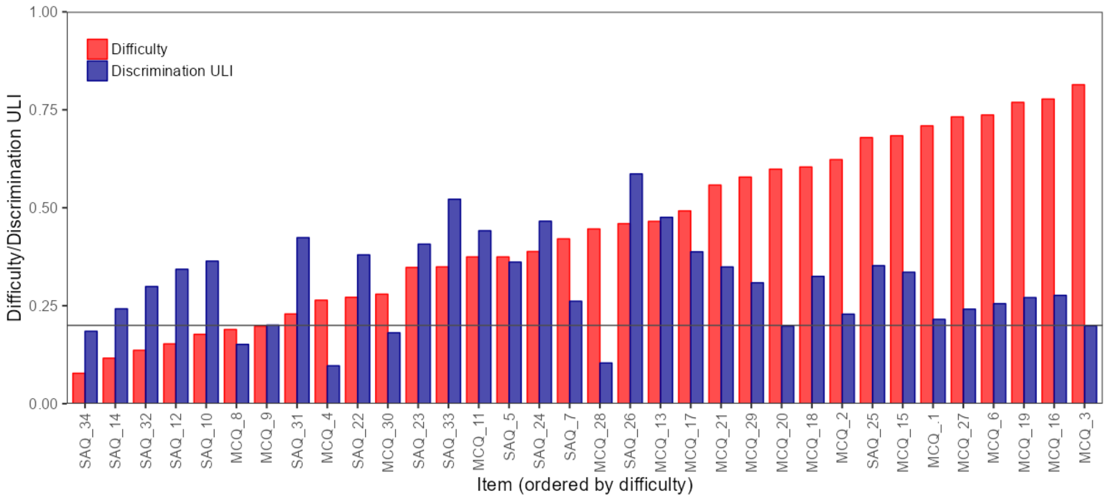

Figure 3 *Item Difficulty and Discrimination Index overview.*

Item Difficulty vs. Discrimination Index

When analyzing item statistics, it is also important to analyze the pattern of the correlation between item difficulty and discrimination index. Even though this illustration tells just limited information about the item's behavior, it supports the analysis (Metsämuuronen, 2018).

When analyzing and deciding about item removal, the correlation between item difficulty and discrimination index should ideally show that items with low discrimination indices are also of inappropriate difficulty. Specifically, items that are either too easy or too difficult should exhibit low discrimination, indicating they do not effectively differentiate between high- and low-ability students. If an item is easy but has high discrimination, it may still be useful; conversely, if it is difficult and also shows low discrimination, it should be considered for removal. While the scatterplot reveals instances where more difficult items have lower discrimination, which may raise concerns about their effectiveness in distinguishing between test-takers of different ability levels, this pattern is not uniform. In fact, the presence of a seemingly random distribution suggests that item difficulty and discrimination may not be strongly or consistently related in our dataset (Upadhyah *et al.*, 2019). This observation aligns with prior research indicating that the relationship between difficulty and discrimination is often non-linear, potentially quadratic, where items of moderate difficulty tend to show higher discrimination (e.g., Fan, 1998).

In many studies, we also observe outliers or items with unusually high or low discrimination given their difficulty level that may warrant further investigation to understand their contribution to the assessment's overall quality and validity. In an ideal scenario (Oermann & Gaberson, 2021), we are looking at items with moderate difficulty levels (neither too easy nor too difficult, between 0.3 to 0.7 range) and high discrimination values (above 0.2), indicating that they effectively differentiate between individuals with varying levels of ability (grey region in Figure 4). Items with high difficulty but low discrimination may need revision or removal from the assessment to improve its psychometric properties (red region in the graph) (Metsämuuronen, 2022).

As a measure of discrimination, we have used Point-Biserial correlation, which is the Pearson correlation between responses to a particular item and scores on the total test (with or without that item). The Biserial Correlation models the responses to the item to represent stratification of a normal distribution and computes the correlation accordingly. The range is -1.0 to 1.0, and generally, 0.2 and above is considered to have high correlation and positive association with overall performance on the assessment.

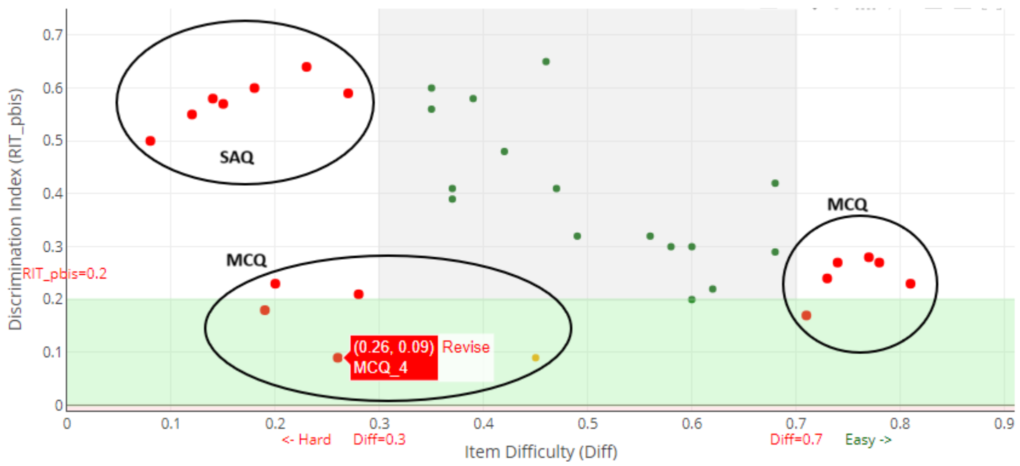

Figure 4. *Item difficulty and discrimination index correlation.*

Figure 4 displays the distribution of the items in relation to item difficulty and discrimination based on empirical data. To aid interpretation, three regions were highlighted in the scatterplot. The upper-left region contained seven SAQ items with discrimination indices above 0.20 and item difficulties below 0.30, indicating challenging items that still discriminated moderately well. The bottom-right region contained only MCQ items (six items), with item difficulties above 0.70 and discrimination indices ranging from 0.17 to 0.28, reflecting relatively easy items with limited discriminative power. The bottom-left region included five MCQ items: four with item difficulties below 0.30 and discrimination indices between 0.09 and 0.23, and one with an item difficulty of 0.45 and discrimination of 0.09. All other items clustered centrally, with item difficulties between 0.30 and 0.70 and discrimination indices above 0.20; this group comprised a mixture of MCQ and SAQ items that generally met acceptable quality standards. This visual makes it easy for subject-matter experts to quickly identify items that performed well or that might need to be revised or removed.

DISCUSSION and CONCLUSION

In Albania, high school exams have been administered in a traditional pen and pencil format since 2006. These exams are important to assess high school students' achievements for university entrance. These exams include mandatory and elective subjects, where each of the two has a combination of marks from multiple-choice questions (MCQ) and open-response questions (SAQ). While the exams rely on expert judgment to classify the difficulty of the questions, the use of psychometric analysis is limited. This study highlights the need for improved methods, highlights the importance of integrating psychometric approaches and historical data to improve question classification and decision-making in future assessments, especially as Albania moves towards electronic assessment practices.

Our findings suggest that while expert judgment is highly valuable in the pre-administration phase of exam development, psychometric analysis after administration provides the statistical evidence necessary to support reliable item classification and future reuse in an item bank. After performing a psychometric analysis of the items and evaluating the reliability of the test for one of the elective subjects, we found that only 44% of the items were classified in the same item difficulty category after the test was administered. This divergence highlights the importance of supplementing expert judgment with more rigorous psychometric and statistical analyses. We highly recommend using previous exam results over the years to support decision-making during the exam development and grading of new test items.

Additionally, based on the psychometric statistics of the MCQ and SAQ items, we recommend rearranging them within the test according to their corresponding topics and subdomains, in order to reduce unnecessary switching between item types and minimize potential student confusion. The creation of a bank of items, as well as psychometric data, will facilitate the development of future examinations and help to more effectively implement electronic assessment methods in the short term.

While our findings suggest notable differences, particularly between how item difficulty is perceived by experts and how it manifests in student performance, we acknowledge that these findings alone are not sufficient to conclude that expert judgment is inadequate.

Follow-up investigations will need to incorporate a broader range of contextual variables to enhance the validity of such findings. Such as: longitudinal, assessment formats, content validity, and expert rationale behind item design. The collaboration between psychometric analysis and expert judgment fosters a robust examination system, but it is vital to adapt and integrate technological advancements to enhance the reliability and effectiveness of assessments.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Eralda Gjika: Conceptualization, methodology, supervision, validation, data analysis, software development, and original draft preparation. **Lule Basha:** Methodology, validation, data analysis, and critical review of the manuscript. **Afërdita Alizoti:** Resources, manuscript review, and interpretation of results. **João Paulo Araujo Lessa**: Review of the manuscript and contributions to methodology.

Orcid

Eralda Gjika https://orcid.org/0000-0003-2662-4316
Lule Basha https://orcid.org/0000-0003-3790-601X
Afërdita Alizoti https://orcid.org/0009-0005-2952-8460
João Paulo Araujo Lessa https://orcid.org/0000-0003-0751-7662

REFERENCES

Bjorner, J.B., Chang, C.H., Thissen, D., & Reeve, B.B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Qual Life Res 16*(Suppl 1), 95-108. https://doi.org/10.1007/s11136-007-9168-6

Brown, G.T.L. (2022). The past, present, and future of educational assessment: A transdisciplinary perspective. *Frontiers in Education*, 7. https://doi.org/10.3389/feduc.2022.1060633

Brownstein, N.C., Louis, T.A., O'Hagan, A., & Pendergast, J. (2019). The role of expert judgment in Statistical Inference and evidence-based decision-making. *The American Statistician, 73*(1), 56-68. https://doi.org/10.1080/00031305.2018.1529623

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Ebel, R.L. (1972). *Essentials of Educational Measurement* (1st ed.). Prentice Hall.

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

Gjika, E., Basha, L., & Alizoti, A. (2025). TIA: A smart psychometric analysis app for Albanian high school exams framework. *Smart Cities and Regional Development* (SCRD) Preprints, 2(1). https://scrd.eu/index.php/scrd-pp/article/view/641

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381. https://www2.hawaii.edu/~daniel/irtctt.pdf

Firescu, A.M., Necula, M., Băsu, C.E., Ardelean, A., Roşu, M.M., Milea, E.C., & Păun, M. (2022). Investigation of the 2019-2021 National Evaluation exam: A case study of high school admissions in Bucharest. *Proceedings of the International Conference on Business Excellence, 16*(1), 59–81. https://doi.org/10.2478/picbe-2022-0008

Ke, C., Yingwei, W., Yajun, Y., & Runhua, L. (2010). *Item-bank in language E-Assessment: Issues and perspectives*. 5th International Conference on Computer Science & Education. Hefei, China. pp. 1657-1660. https://doi.org/10.1109/ICCSE.2010.5593611

Kreitchmann, R.S., Abad, F.J., Ponsoda, V., Nieto, M.D., & Morillo, D. (2019). Controlling for Response Biases in Self-Report Scales: Forced-Choice vs. Psychometric Modeling of Likert Items. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02309

Maghnouj, S., et al. (2020). OECD Reviews of Evaluation and Assessment in Education: *Albania, OECD Reviews of Evaluation and Assessment in Education*. OECD Publishing, Paris, https://doi.org/10.1787/d267dc93-en

McAlpine M, Hesketh, I. (2003). *Multiple response questions- allowing for chance in authentic assessments*. In: Christie J, editor. 7th International CAA Conference. Loughborough University.

Metsämuuronen, J. (2018). Essentials of visual diagnosis of test items – Logical and pathological patterns in items to be detected. https://doi.org/10.13140/RG.2.2.13950.23364

Metsämuuronen, D. (2022). Essentials of Visual Diagnosis of Test Items. Logical, Illogical, and Anomalous Patterns in Tests Items to be Detected. *Practical Assessment, Research, and Evaluation, 27*(1), 5. https://doi.org/10.7275/n0kf-ah40

MoES. (2019). Matura shtetërore: Gjuha shqipe dhe letërsia [Exam guideline]. https://maturashteterore.com/wp-content/uploads/2020/03/gjuhe_shqipe.pdf

MoES. (2023). Urdhër i përbashkët për Maturën Shtetërore [Regulatory document]. https://arsimi.gov.al/wp-content/uploads/2025/05/URDHER-I-PERBASHKET.pdf

OECD (2024a, June 18). PISA 2022 Results (Volume III): Creative minds, creative schools. OECD Publishing. https://doi.org/10.1787/765ee8c2-en

OECD (2024b, March 1). PISA 2022 Technical Report, PISA. OECD Publishing. https://doi.org/10.1787/01820d6d-en.

Oermann, M.H., & Gaberson, K. (2021). Evaluation and testing in nursing education. Springer Publishing.

Revelle, W. (2024). psych: Procedures for psychological, psychometric, and personality research (Version 2.3.12) [R package manual]. https://cran.r-project.org/web/packages/psych/psych.pdf

Saito T, Rehmsmeier M (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Schwarz, R., Bulut, H.C., & Anifowose, C. (2023). A data pipeline for e-large-scale assessments: Better automation, quality assurance, and efficiency. *International Journal of Assessment Tools in Education, 10*(Special Issue), https://doi.org/10.21449/ijate.1321061

Schwarz, R., & Gjika, E. (2023). *Modernized psychometric analysis for digital assessments: An open-source approach for automation, quality assurance, and efficiency*. The 48th International Association for Educational Assessment Annual Conference. https://iaea2023.org/

Shahini, A. (2021). Inequalities in Albanian education: Evidence from large-scale assessment studies. *Kultura i Edukacja*, *134*(4), 40–70. https://doi.org/10.15804/kie.2021.04.03

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Suto, I., Crisp, V., & Greatorex, J. (2008). Investigating the judgemental marking process: an overview of our recent research. https://doi.org/10.17863/CAM.100464

Ten Berge, J.M.F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613-625. https://doi.org/10.1007/bf02289858

Upadhyah, A.A., Maheria, P.B., & Patel, J.R. (2019). Analysis of one best MCQS in five preuniversity physiology examinations. *International Journal of Physiology*, *7*(4), 10-15. https://doi.org/10.37506/ijop.v7i4.47

Warburton, W.I. & Conole, G.C. (2003). Key Findings from recent literature on Computer-aided Assessment. *ALT-C 2003, Sheffield. 08-10 Sep 2003.* 19 pp. https://eprints.soton.ac.uk/14113/

Wonde, S.G., Tadesse, T., Moges, B., & Schauber, S.K. (2024). Experts' prediction of item difficulty of multiple-choice questions in the Ethiopian undergraduate medicine licensure examination. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-06012-x

Wright, B.D., & Bell, S.R. (1984). Item Banks: What, Why, How. *Journal of Educational Measurement*, *21*(4), 331–345. http://www.jstor.org/stable/1434585