EXPLORING THE DYNAMICS OF ASSESSMENTS: A COMPREHENSIVE REVIEW OF MCQs, SAQs, AND E-ASSESSMENT IN ALBANIA

ERALDA GJIKA (DHAMO)[1], LULE BASHA (HALLAÇI)[2], AFËRDITA ALIZOTI (MËHILLI)[3]

[1,2]Department off Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania

[3]Centre of Educational Services, Tirana, Albania

e-mail: eraldagjika@gmail.com

Abstract

This paper advocates for a transformative shift in assessment methodologies by integrating Multiple-Choice Questions (MCQs) and Short Answer Questions (SAQs), considering optimal sample sizes, and adopting e-assessment platforms. It emphasizes surpassing minimum sample size requirements for robust psychometric analyses and highlights the flexibility of e-assessment platforms for enhanced precision. Integrating MCQs and SAQs within an e-assessment framework equips item developers with comprehensive information for improvements and empowers psychometricians with advanced analysis capabilities. The discussion extends to challenges and advancements in implementing these question formats in Albanian state exams, emphasizing the need for enhanced competence in test development and thorough psychometric analysis.

Key words: *MCQ, SAQ, E-assessment, psychometric analysis, exam Reliability.*

Përmbledhje

Ky punim promovon idenë e një ndryshimi transformues në metodologjitë e vlerësimit duke integruar pyetje me alternativa (MCQs) dhe pyetje me përgjigje të shkurta (SAQs). Ai merr në konsideratë madhësitë optimale të zgjedhjes dhe vë theksin në rolin e platformave të vlerësimit elektronik. Ai thekson kërkesat mbi vëllimin minimal të zgjedhjes për analiza psikometrike të qëndrueshme dhe njëherëshi thekson fleksibilitetin e platformave të vlerësimit elektronik për të përmirësuar cilësinë e analizave. Integrimi i MCQs dhe SAQs në një strukturë e-vlerësuese pajis zhvilluesit e pyetjeve me informacion të plotë për përmirësime dhe fuqizon psikometritkanët duke i *dhënë aftësi të përparuara në analizën e testeve. Diskutimi në këtë punim zgjerohet në sfidat dhe përparësitë në implementimin e këtyre formateve të pyetjeve në provimet shtetërore në Shqipëri, duke theksuar nevojën për kompetencë të përmirësuar në zhvillimet e testeve dhe analizat psikometrike të tyre.*

Fjalë kyçe: *Pyetje me Alternativa, pyetje me përgjigje të shkurta, vlerësim elektronik, analiza psikometrike, besueshmëria e testit.*

Introduction

The significance of e-assessment in modern education lies in its transformative impact on the learning landscape, offering efficiency, scalability, and adaptability. Well-crafted questions play a critical role as they serve as the cornerstone of effective assessments. These questions in most cases are carefully designed to align with educational objectives, ensuring they not only measure knowledge but also promote deeper understanding and critical thinking. The emphasis on well-crafted questions underscores the need for thoughtful and purposeful design to maximize the educational value derived from assessments, making them powerful tools for enhancing learning outcomes. In the digital era, e-assessment provides educators with tools to tailor assessments to diverse learning styles, track individual progress, and offer timely feedback.

Within the domain of e-assessment, Multiple-Choice Questions (MCQs otherwise known as SBA- Single Best Answer) and Short Answer Questions (SAQs) emerge as two widely adopted formats, each carrying distinct advantages and disadvantages. Exploring these question types within the evaluation of e-assessment allows for an in-depth understanding of their effectiveness, challenges, and unique contributions to evaluating learners' knowledge and skills. MCQs, renowned for their efficiency, allow rapid, automated grading, ensuring standardized assessments with broad content coverage. However, they may fall short in evaluating deep understanding and critical thinking, as they often focus on recall rather than application. On the other hand, SAQs offer a more nuanced approach, assessing comprehension and application of knowledge.

They provide flexibility in question types and reduce the impact of lucky guesses, yet their manual grading can be time-consuming and subjective. The choice between MCQs and SAQs hinges on educational objectives, desired depth of understanding, and logistical considerations, with a strategic combination of both formats often proving beneficial in achieving a comprehensive assessment. Due to recent advances in technology, they both can be delivered, marked electronically, and obtain the item performance in a few minutes. In their paper (Schwarz et al., 2023) present a user-friendly modernized interface of the current psychometric analysis practices. They share details and further improvements on the workflow design and functionalities of data pipeline, suggesting a forward-looking and innovative exploration of psychometric analysis within the context of digital assessments. Their demonstration of how their pipeline works as an accurate tool to calculate psychometric results within minutes highlights the efficiency and reliability of their approach (Schwarz & Gjika, 2023). The open-source approach is a commitment to transparency and accessibility in psychometric analysis tools or methods, potentially promoting collaboration and community involvement. This aligns with the broader trend in academia and technology towards open-source solutions.

In the context of digital assessments, the decision to use MCQs or SAQs can be strategically aligned with the principles of automation, quality assurance, and efficiency in psychometric analysis. Automation, with its focus on a streamlined and faster analysis process, aligns well with the objective nature of MCQs, where answers can be automatically graded. The predefined answer options and clear-cut scoring criteria contribute to a more objective evaluation, enhancing the reliability of assessments. While MCQs align well with the principles of automation, quality assurance, and efficiency, it is important to acknowledge that SAQs, with their focus on deeper understanding and application, may have unique merits in certain educational contexts. The specific learning objectives, the depth of understanding to be measured, and the overall goals of the assessment should inform the choice between MCQs and SAQs.

In their study (Mee et al., 2023) compared the difficulty, discrimination, and time requirements for the two formats when examinees responded as part of a large-scale, high-stakes medical assessment by converting a considerable amount of MCQs items to SAQs. Their study showed that items administered in the SAQ format were generally more difficult than items in the MCQ format. The item difficulty, discrimination index and response time of administration increased significantly when passing from MCQ to SAQ. The same observations were previously confirmed by (Sam et al., 2019) applied in medical in large-scale, high-stake assessments. The combination of MCQs and SAQs is commonly employed in medical assessments due to the diverse learning objectives and complex nature of medical education. Medical assessments adopt a comprehensive approach to evaluating medical students, considering not only their foundational knowledge but also their ability to apply this knowledge across various clinical scenarios. This combination ensures that assessments align with the multifaceted nature of medical education and adequately prepare students for the complexities of medical practice. Recent developments in automated scoring systems have rendered the substitution of MCQs with SAQs feasible in extensive, consequential evaluations.

Instead, in primary education, where the focus often lies on assessing proficiency levels with a pass or fail outcome, the predominant use of MCQs may seem more practical. MCQs offer a straightforward and efficient way to assess foundational knowledge across various subjects. Their objective nature makes them easier to administer, grade, and standardize, especially when dealing with large cohorts of students. However, it is essential to acknowledge that while MCQs can effectively measure recall and recognition skills, they may not fully capture students' deeper understanding of critical thinking abilities. Therefore, incorporating a mix of question formats, including SAQs, could enrich the assessment process and provide a more holistic view of students' learning. In Albania, the primary education system commonly emphasizes evaluating proficiency levels through examinations like the grade 5 exam (alb. VANAF-Vlerësimi i Arritjeve të Nxënësve të Arsimit Fillor). This assessment serves to measure student and teacher performance within the elementary program curriculum of three primary subjects: Albanian Language, Mathematics, and Natural Science. While not mandatory, it plays a significant role in measuring student achievement and guiding instructional strategies. It is presented in a paper format, incorporating both MCQ and SAQ offering a practical means of evaluating foundational knowledge in diverse subjects.

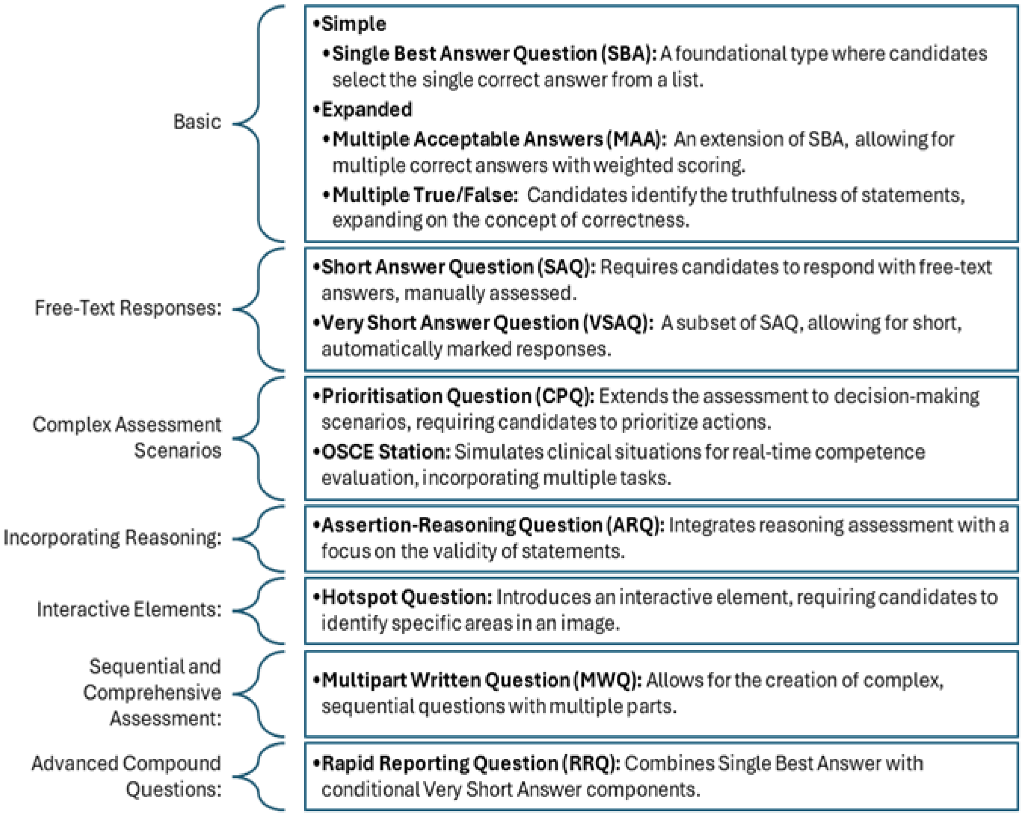
In contrast, high school education exams in Albania serve as a pivotal step towards university entrance, necessitating a more comprehensive evaluation of students' knowledge and skills. Here, the use of SAQs becomes increasingly valuable. SAQs allow students to demonstrate not only their factual knowledge but also their ability to analyze, synthesize, and apply concepts in various contexts. This aligns more closely with the expectations of higher education institutions and the demands of modern workplaces. Additionally, SAQs encourage students to develop critical thinking, problem-solving, and communication skills, which are essential for success beyond high school.

While the transition from MCQs to SAQs may require additional resources and training for educators, the benefits in terms of assessing higher-order thinking skills and preparing students for university-level studies are significant.

This approach allows for a targeted exploration of the challenges and considerations associated with these question types, contributing to a more nuanced understanding of effective assessment practices.

MCQ and SAQ Guidelines for Item Development

Understanding the nuances of various item types allows for a comprehensive exploration of communication skills and reasoning abilities within a diverse range of assessment scenarios. The conceptual flow of item types, ranging from foundational to advanced types, build a foundation of this exploration, offering a structured framework to understand their interrelationships and potential impact on communication competence. Below, we present a conceptual flow illustrating the categories and possible relationships among some of the most common item types, Figure 1.



Basic

- **Simple**
  - **Single Best Answer Question (SBA):** A foundational type where candidates select the single correct answer from a list.
- **Expanded**
  - **Multiple Acceptable Answers (MAA):** An extension of SBA, allowing for multiple correct answers with weighted scoring.
  - **Multiple True/False:** Candidates identify the truthfulness of statements, expanding on the concept of correctness.

Free-Text Responses:

- **Short Answer Question (SAQ):** Requires candidates to respond with free-text answers, manually assessed.
- **Very Short Answer Question (VSAQ):** A subset of SAQ, allowing for short, automatically marked responses.

Complex Assessment Scenarios

- **Prioritisation Question (CPQ):** Extends the assessment to decision-making scenarios, requiring candidates to prioritize actions.
- **OSCE Station:** Simulates clinical situations for real-time competence evaluation, incorporating multiple tasks.

Incorporating Reasoning:

- **Assertion-Reasoning Question (ARQ):** Integrates reasoning assessment with a focus on the validity of statements.

Interactive Elements:

- **Hotspot Question:** Introduces an interactive element, requiring candidates to identify specific areas in an image.

Sequential and Comprehensive Assessment:

- **Multipart Written Question (MWQ):** Allows for the creation of complex, sequential questions with multiple parts.

Advanced Compound Questions:

- **Rapid Reporting Question (RRQ):** Combines Single Best Answer with conditional Very Short Answer components.

Figure 1. Common Item Types used in Educational Tests

The above figure illustrates how certain item types serve as foundational or expanded versions of others and can be applied to diverse assessment needs and complexity levels. However, this relationship may vary based on specific assessment goals and contexts. In this study, our focus will be directed towards two main and useful item types: Multiple-Choice Questions (MCQ, or otherwise known as Single Best Answer Question -SBA) and Short Answer Questions (SAQ).

While constructing MCQs, some essential guidelines should be considered to ensure the effectiveness of the assessment (Al-Rukban, 2006). The anatomy of a MCQ comprises several essential elements. Starting with clarity in the wording of the source and response choices are essential. It is important that each question should be clearly aligned with the learning objectives, ensuring that the assessment accurately measures the target knowledge or skills. Correct answer (key) and distractors must be homogenous and carefully crafted to assess targeted knowledge or skills. The key serves as the definitive correct response, while distractors aim to identify common misconceptions. The item length and structure should be consistent to avoid unintended cues to the correct answer and the use of negative or double negative phrases should be avoided to prevent confusion. In some specific cases, feedback may be provided to enhance the educational value of the assessment. The inclusion of context, images, or diagrams in MCQs is particularly relevant in the context of the evolution of e-assessment and the tools that support it. As digital assessment platforms have advanced, they have provided the capability to integrate multimedia elements seamlessly into questions (Schwartz et al., 2023; Schwartz & Gjika, 2023). One of the main advantages is that it allows for the presentation of complex scenarios, clinical cases, or real-world contexts, mirroring the challenges professionals might encounter in their fields. Apart from making assessments more engaging, it also fosters a more authentic evaluation of a test-taker's ability to apply knowledge in practical situations. The visual elements addition will not always aid in assessing visual literacy and interpretation skills or offering a more comprehensive evaluation beyond traditional text-based questions as demonstrated by (Holland et al., 2015).

On the other hand, SAQs should be clearly formulated, avoiding ambiguity or multiple interpretations. A clear and concise stem, along with precise instructions, sets the context and guides test-takers on the expected response format. They should align closely with the learning objectives, emphasizing higher-order cognitive skills such as analysis, synthesis, and application. A well-defined scoring rubric promotes objectivity in evaluation. Ensuring clarity and uniqueness in questions avoids misinterpretation, and relevance to real-world scenarios enhances the application of knowledge. SAQs should be designed to assess specific and targeted knowledge or skills, avoiding overly broad inquiries.

When creating SAQs, educators often align the questions with specific levels of Bloom's Taxonomy to target the desired cognitive skills (Bloom et al., 1956). Other taxonomies or frameworks related to cognitive skills may also be used depending on educational goals and preferences. In their study (Nguyentan et al., 2022) developed a checklist testing with pharmacy students at the University of California San Francisco that they suggested may be used by faculty members when developing SAQs. This was previously suggested in other similar studies (Przymuszała et al., 2020; Hijji, 2017; Al-Rukban, 2006) which indicated that distributing a concise guideline document among academic teachers can effectively reduce the number of item-writing flaws (IWFs) in multiple-choice questions (MCQs). These flaws can include ambiguities, inconsistencies, or biases in the wording of the question or its answer choices, which may affect the validity and reliability of the assessment. Other researches (Raymond et al., 2019) have shed light on the optimal format for multiple-choice questions (MCQs), emphasizing the importance of the three-option format. These findings highlight the importance of reevaluating traditional methodologies in item and test development to ensure the validity and effectiveness of educational evaluations.

Field trial importance in education assessment

The culture of testing, through processes like piloting and field trials, plays a pivotal role in the development of MCQ and SAQs items. Testing is indispensable as it allows educators and assessment developers to evaluate the effectiveness, clarity, and appropriateness of questions before deploying them on a larger scale. It provides valuable insights into how test-takers interpret and respond to the questions, helping identify potential ambiguities, misconceptions, or biases. This way we ensure the reliability and validity of assessments and allows us for refinement based on empirical evidence, contributing to the overall quality of the assessment instrument.

According to Boruch, 2003, randomized field trials (RFTs) play a vital role in educational research, offering a systematic approach to assessing the effectiveness of educational interventions. In these trials, participants are randomly assigned to different groups receiving various educational interventions. One key benefit of RFTs is their ability to facilitate fair comparisons between different regimens, ensuring that estimates of outcome differences are statistically unbiased. Additionally, RFTs enable researchers to make confident statistical statements about the results, accounting for variability in institutional and human behavior. By randomly assigning entire institutions or jurisdictions to different regimens, RFTs help estimate the effects of interventions on a broader scale, providing valuable insights into educational practices.

While less common in education, RFTs have been conducted at various levels and in different countries, contributing to advancements in educational research and practice. The Field Trial sample size for assessed students in PISA 2025 technical standards (PISA, 2025) is emphasized and it is determined by the test design and language of assessment. It aims to achieve 200 student responses per item in the largest language of assessment within each adjudicated entity. Additionally, for other assessment languages covering at least 5% of the target population, a minimum of 100 students per item is required. The same minimum sample size applies to additional adjudicated entities with assessment languages used by at least 5% of their target population.

Sample size and its impact on psychometric

Understanding and appropriately addressing sample size is paramount in psychometric research as it directly influences the reliability and generalizability of study findings. The significance of sample size lies in its ability to impact the precision and statistical power of analyses, ultimately influencing the validity of study conclusions. In the subsequent review attention is directed towards illuminating insights and practical tips when confronted with low sample sizes, acknowledging the challenges posed and proposing strategies to enhance the methodological robustness of psychometric research in such instances.

Statistical power directly relates to the ability of a study to detect a true effect or difference when it exists. In essence, it quantifies the likelihood that a study will correctly reject the null hypothesis when the alternative hypothesis is true. A low statistical power may fail to detect real effects, leading to false conclusions or missed opportunities to identify important findings in a study. One can imagine statistical power as the focus capability of a camera lens. Much like a camera lens needs to be appropriately adjusted to capture fine details in a photograph, statistical power in a research study is essential for detecting subtle effects. If you set a low focus on the camera lens, you might miss out on intricate elements in your photo. Similarly, in a study with low statistical power, there's a higher chance of overlooking nuanced effects. Just as a well-adjusted lens enhances the clarity of an image, a study with sufficient statistical power improves the likelihood of detecting real effects. This analogy underscores the importance of choosing the right "focus" or statistical power to ensure the study captures meaningful insights (Coolican, 2014).

Therefore, determining an appropriate sample size based on the desired level of statistical power is essential for ensuring the reliability and validity of study results (Kyriazos, 2018). Four crucial parameters are considered in power analysis of a study(Barker et al., 2015): 1) Sample size (N); 2) Alpha, representing the probability of identifying a non-existing effect (Type I error), the standard cutoff generally is set .05; 3) Beta, representing the probability of not identifying an existing effect (Type II error), with statistical power calculated as (1 − Beta), generally a power of .80 is desired; 4) Effect size, indicating the strength of the examined relationship, categorized as small, medium, or large. Considering these factors during study planning is crucial for determining the appropriate sample size, and omitting this step could lead to the potential failure to detect significant effects (Kyriazos, 2018). In their article (Morgado, et al., 2017) systematically review scale development practices in human and social sciences, covering 105 studies from 1976 to 2015. It identifies ten main limitations, including issues with sample characteristics, methodology, psychometrics, qualitative research, missing data, social desirability bias, item constraints, brevity of scales, difficulty controlling variables, and absence of manual instructions. While recognizing methodological weaknesses, especially in psychometric analysis with smaller sample sizes, the article emphasizes the importance of acknowledging and addressing these limitations for improved future scale development research practices.

Building upon this research (Hackshaw, 2008) highlights the need for a balanced approach between small studies, which can be conducted quickly, and larger studies, which may take several years to complete. It emphasizes the importance of defining what is considered "small" based on study objectives and discusses the challenges and considerations associated with different sample sizes, particularly in the context of clinical research studies. The article also underscores the strengths of small studies, including their quick execution, ease of ethical and institutional approval, and the use of surrogate markers to expedite observations.

Furthermore, in a related study (Faber & Fonseca, 2014) explores the importance of sample size calculation in epidemiological, clinical, and lab studies, highlighting the ethical and methodological considerations. It emphasizes the potential impact of varied sample sizes on clinical decision-making, advocating for a nuanced approach to ensure meaningful study interpretation. Moreover (Ahrens & Zaščerinska, 2014) highlights the dual role of sample size in research, emphasizing its connection to both statistical analysis and generalization. It underscores a gap in attention to a framework for selecting sample size in educational research on E-Business application.

Employing explorative and interpretive research paradigms, the study identifies key concepts related to sample size and proposes a framework based on findings from four expert perspectives, outlining directions for future research. Extending the discussion further (Vasileiou et al., 2018) underscores the conceptual and practical significance of selecting an appropriate sample size in qualitative research. Analyzing single-interview-per-participant designs in health-related journals, the findings reveal limited justification for sample sizes, often characterized as insufficient and discussed in the context of study limitations, threatening the validity and generalizability of results.

Complementing these insights (Besekar, 2023) conducted an extensive review to evaluate sample size considerations in educational research, recognizing the critical role of specifying the number of participants representing the target population. The review extracted information from databases such as Google Scholar and PubMed, leading to the selection of seven articles. The findings emphasized the variability in sample size determination methods across different study designs and concluded that, on average, 24.24 participants per group were required for testing novel approaches, with a median sample size of 30 for simulation-based educational research. The study concludes that considering sample size early in the research phase is crucial, with the selection of an appropriate formula depending on various factors such as the study's goal, outcome variable, plan, statistical investigation, and other logistical considerations.

While larger samples enable detailed analysis and ensure stability in item calibrations, small sample studies in educational research also hold value. Despite their limitations, small samples can yield valuable insights, especially in studies employing techniques like role-playing and in-depth interviews. Borg (1987) emphasizes the achievement of validity, reliability, and generalizability in small sample studies, albeit through a redefined understanding of these terms. In educational research, where studying entire populations is often impractical, small samples can be more appropriate, offering nuanced knowledge that may be overlooked in larger-scale studies (Borg & Gall, 1989). Even in medicine, single-case investigations provide valuable insights into treatment approaches. Therefore, while large samples offer advantages, small sample studies remain crucial contributors to educational research.

Addressing the challenge of small sample sizes (Wright & Stone, 1979), researchers have shown that valuable insights can still be gleaned from samples as small as 35 students and 18 items. While small sample sizes pose limitations across statistical methodologies, simulations based on estimates can enhance result credibility. In justifying a low sample size for Rasch analysis (where the minimum participants required is 30), it's crucial to highlight the method's specific requirements and emphasize the societal benefits of contributing to the field, even with smaller samples.

According to the findings of (Nunnally, 1978) review study, they underscore the importance of future research endeavors aiming for a larger sample size, with a suggested minimum ratio of 10:1 (N:p) as a rule of thumb ratio of cases to free parameters, advocating for ten participants for each variable or parameter being analyzed. This proportion is advised to enhance the statistical power, reliability, and generalizability of research outcomes. In psychometric analysis, it proposes having a sufficient number of participants compared to the intricacy of the variables being examined, ensuring a robust and reliable evaluation of the phenomena being investigated. The next section gives an overview of sample size possible limitations in three main state exams in Albania.

Case Study: Implementation and Impact of Exam Reforms in Albania

In this section, we undertake a comprehensive exploration of three pivotal examinations in Albania. The State Matura exams, Regulated Professions exams; University Entry exams for public universities (Medical University of Tirana).

By exploring the historical origins to gain insights into the evolution of these assessments and understanding their necessity, we expose the societal and educational context. Examining how these exams are conducted then illuminates the mechanisms employed to ensure fairness. Additionally, a closer look at the challenges faced offers a nuanced perspective on the continuous efforts to refine and optimize these critical assessment processes. Our focus will be on understanding the historical origins, test formats, and trends in participant numbers over the years. While we won't delve into detailed test performance analysis, our objective is to offer insights into best practices based on existing studies and observations. Through this exploration, we aim to inform policymakers and stakeholders about potential areas for improvement and recommend strategies for enhancing the effectiveness and fairness of these assessments.

For all three above-mentioned exams, the Center for Educational Services (CES) plays a central regulatory role throughout the examination process, overseeing test development, exam creation, marking procedures, performance evaluation, and the delivery of final student results. More specifically the role of CES for the State Matura exam, is to manage the entire process from the design of the item pool used for the final test, evaluation scheme design, evaluation process, and announcing the students' results and excluding test administration. Meanwhile, for regulated professions and the University entry exams, CES collaborates with HEI's (Higher Education Institutions) professors for drafting and editing the item pool, generating the exams, administering the testing process, and announcing the final results, all conducted via computer-based tests (CES, Regulated Profession Item Fund, 2023; CES, State Matura item Fund, 2023).

i) State Matura Exams

Genesis and Inception: The Commencement of Educational Reforms

The introduction of the State Matura in Albania in 2006 marked a significant shift in the examination system for pre-university education. The primary objective was to revamp the assessment process, with State Matura Exams serving a dual purpose: acting as the final evaluation for pre-university education and serving as the selection criteria for admission to public higher education institutions. Prior to 2006, matriculation exams were solely the concluding step in pre-university education, and entrance into public universities relied on institution-specific entrance exams. The matriculation exams were previously crafted by the Ministry of Education and assessed internally by subject-specific teachers in each school. This decentralized approach led to a lack of standardization, allowing different teachers to employ varied criteria in assessing student responses. The implementation of the State Matura reform brought about comprehensive standardization in the entire process. This encompassed the establishment of exam centers, the testing environment, form completion, test design, assessment of student responses, and the issuance of the State Matura diploma. Although the foundational principles of the State Matura reform remained unchanged, the system underwent continuous modifications over the years.

The State Matura exams are administered in a traditional pencil and paper format. Students specializing in general high school profiles are required to undergo three mandatory exams along with one optional exam. The obligatory exams cover the following subjects: Foreign language (English, French, German, Italian, Spanish, and Turkish); Albanian language and literature; Mathematics (State Matura regulation, 2022). Additionally, students have the flexibility to select one optional subject from the list of eight subjects provided in the table below.

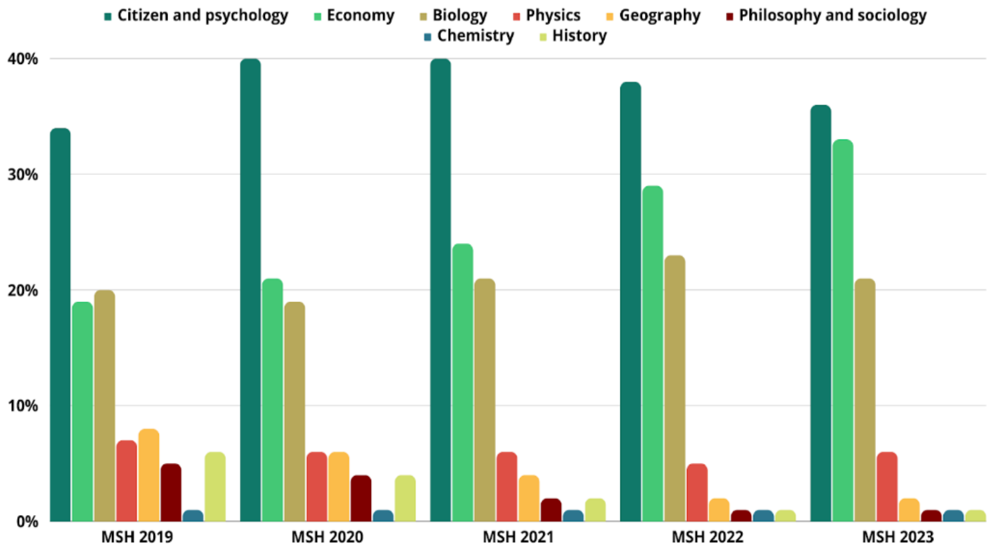


Figure 2. The distribution of optional subjects chosen by students in State Matura (period: 2019-2023; alb. MSH-Matura Shtetërore eng.State Matura)

Figure 2 illustrates the percentage of graduates opting for each optional subject between 2019 and 2023. This timeframe aligns with the integration of the competency-based curriculum into the State Matura. The graph spanning from 2019 to 2023 reveals notable trends in student subject selection across eight disciplines. Notably, Economy demonstrates a consistent upward trajectory, with an increase of 2% observed annually over the consecutive years. Conversely, subjects such as Geography exhibit a contrasting pattern, experiencing a steady decline of approximately 2% each year. The high school curriculum, unchanged since 2019, has different weights allocation of hours for optional subjects. Economics receives only 3 hours per week and is taught only in the third year, while Citizenship and Psychology are limited to 2 hours each in two different years and are popular choices due to a small weight of topics taught. Conversely, Biology is particularly popular due to its relevance to medical fields. This trend underscores the dynamic nature of student preferences and highlights the importance of monitoring and adapting educational topics to align with evolving interests and needs.

Diverse Challenges and Strategic Choices: Navigating Exam Variances and Subject Selection Patterns

Examining the graph reveals a distinct preference among graduate's subjects, particularly citizenship-psychology and economics. Notably, these subjects carry a considerably lighter teaching load, measured in teaching hours, compared to subjects like physics, chemistry, history, and geography. The latter subjects demand 2 hours in the 10th grade, 2 hours in the 11th grade, and 4 hours in the 12th grade, unlike the Citizenship-Psychology exam. In this case, citizenship requires 2 hours in the 10th grade, while psychology necessitates 2 hours in the 12th grade. Economics, on the other hand, only requires 3 hours in the 12th grade.

This uneven distribution of teaching hours in the high school curriculum results in disparate weights for each subject, subsequently influencing the choices made by students when selecting exam subjects. Students tend to gravitate towards subjects with a reduced orientation program, finding it more manageable than subjects with heavier workloads. Other factors, including admission criteria for higher education, also play a role in the decision-making process for elective courses. Teachers specializing in relevant high school subjects serve as question makers, developing a question pool constituting no less than 10% of the total test questions for each State Matura test, resulting in a pool of at least 350 questions.

From 2006 to 2017, compulsory subject exams comprised 25 questions (50 scores), with 13 questions (13 scores) featuring as multiple choice and 12 questions (37 scores) requiring open responses. In 2006 and 2007, elective exams were multiple-choice with 50 scores. From 2008 to 2018, the format shifted to include MCQ and open response questions, totaling 40 scores, with 10 scores allocated for alternative questions and 30 scores for open questions (short and/or long and essay), amounting to a total of 20 questions; 10 with alternatives and 10 open. Starting from the State Matura in 2019 and continuing, all tests carry a total of 60 scores, with 20 scores designated for multiple-choice questions and 40 scores for open questions, including structured and essay-type questions in Albanian literacy and Foreign language. The number of open questions varies across exams, generally ranging from 13 to 16.

The Impact of COVID-19 on Exam Dynamics and Student Choices

In the 2020 State Matura, influenced by the pandemic, the test maintained 60 scores and exclusively featured MCQ for all subjects. Answer evaluations were conducted through the scanning of answer sheets. Additionally, the passing threshold has undergone a shift. Between 2006 and 2019, the passing threshold fluctuated annually, contingent on the overall performance of graduates. Conversely, as of 2020, graduates must attain 25 percent of the test scores to achieve passing status (MoES, 2023).

Blueprint and Organizational Framework: Structuring the State Matura Examinations

Since 2006, the design of State Matura tests relies on the subject-specific orientation program, forming the basis for completing the table of test specifications or the test blueprint. The orientation program determines the hours or weight assigned to each topic, subsequently influencing the allocation of scores. Therefore, the subject with the highest weight in the orientation program corresponds to the test section with the greatest number of scores. The State Matura questions exhibit varying levels of difficulty, with 40% classified as basic level, 40% as average level, and 20% as high level. Evaluators employ an orientational evaluation scheme equipped with possible correct answers for each question. Prior to the evaluation process, all certified evaluators undergo comprehensive training for each potential answer to the test questions. Evaluators must have practiced their profession as teachers for at least 5 years, which can then be certified. Subsequently, the schema is made publicly available.

Upon completion of higher secondary education, the graduation process underwent changes over the years. Until 2005, students received a certificate of maturity directly from the school. From 2006 to 2010, the school issued the certificate, which was then signed at the relevant regional educational office. However, starting in 2011, the State Matura diploma, in conjunction with the grade certificate spanning all high school years, became the recognized credential. The Center for Educational Services (CES) issues diplomas, while the school utilizes a unified online interface for printing the grade certificates. The State Matura data system is organized to secure an electronic data archive, mitigating the risk of document falsification related to the successful completion of pre-university education (State Matura regulation, 2024).

ii) *Digital Testing for Regulated Professions in Albania*

Since 2009, various professions in the Republic of Albania, including doctors, dentists, pharmacists, nurses, and others, have been subject to legal regulation (MoES, Regulated Professions in Albania Law, 2009). However, as of now, the licensing procedure for architects, engineers, and social workers among these regulated professions is yet to be fully developed. To qualify for practicing a regulated profession, candidates must engage in professional practice, undergo a state exam, and be registered in the relevant Professional Orders. Professional practice involves acquiring technical, practical, and ethical knowledge under the guidance of a professional in the respective field. Besides exam development for regulated professions, digital tests are conducted for diverse purposes, such as teacher and nurse employment. Since 2023, a computerized test has served as the entrance exam for the general medicine study program at public universities.

The question pool for each regulated profession, containing at least 1500 multiple-choice questions without correct answers, is published on the official website of the CES. Each testing session lasts 60 minutes, featuring 50 questions from the published pool, with a total of 100 scores. The digitized testing format allows for the random generation of tests, ensuring a unique set of questions for each session while maintaining consistency among candidates. Although items may be randomly re-ordered, the alternatives remain in the same sequence for all candidates. It is important to note that multiple sessions may occur during the day, depending on the candidate registration number, and each session comprises different items in their respective tests.

iii) *Computerized University Entry Exam (General Medicine)*

In a groundbreaking initiative in Albania, the University of Medicine, Tirana introduced a computerized entrance exam for the "General Medicine" study program for university admission academic year 2023-2024 (August 2023). This new digital testing method enhances transparency and safety in the public domain while serving as a valuable tool to uphold the quality of university admissions through a meritocratic approach. While its success remains uncertain at this initial stage, the initiative's potential achievements could spark broader adoption by universities aiming to modernize their testing approaches and fortify the integrity of their admissions procedures. The question pool for computerized medicine testing comprises a total of 3000 questions, covering the subjects of biology, physics, and chemistry, each with

1000 questions. The allocation of background questions for each subject is determined based on the specific weight defined in the blueprints or guidelines, ensuring the topics with the highest weight receive the greatest number of questions in the pool.

All background questions are in a multiple-choice format (MCQ), featuring four alternatives, of which only one is correct. The scoring system for these questions varies, with item difficulty levels, decided based on the item developer examination team's experience, determining whether a question is worth 1, 2, or 3 scores. It is noteworthy that none of the items has undergone previous psychometric analysis. To ensure the scientific accuracy of the question bank, the Computerized Testing Committee (2 representatives from the state University of Medicine and 3 high school teachers' representatives from MoES (Ministry of Education and Sport)) meticulously reviewed and approved it. This scrutiny focused on maintaining equal subject representation, assessing question difficulty, and validating the scientific accuracy of the question pool.

Given the resemblance between the regulated professions exams and the medical university admission, the model employed for the former aligns with the structure utilized in medical exams for university admissions. Questions without the correct answers are accessible on the official website of the Educational Services Center. Each testing session spans 60 minutes, featuring 50 questions from the published pool, totaling 100 scores. The digitized format allows for the random generation of unique tests for each session, while all participants share the same test with a different order of questions. The standardized computerized testing structure remains consistent across all sessions, commencing and concluding simultaneously for all candidates on the specified day and time, with a 60-minute duration. Candidates have the flexibility to modify their answers during the test until the time allocated for the test ends automatically.

Automated evaluation by the computer system occurs immediately at the test's conclusion. The individual test results, along with the given answers, are displayed on each candidate's monitor and automatically printed by the system at the end of the test. Noteworthy statistics from the computerized medicine testing include an average State Matura grade of 9.30 and an average test score of 46.46 for all participants (1022 candidates). Among the winners (400 participants), the average State Matura grade rises to 9.61, with an average test score of 63.

Strengths of this process include:

- Enhancing transparency and credibility in the public.

- Real-time announcement of results for candidates.

- Publication of questions to prevent various forms of corruption.

Weaknesses of this process include:

- The absence of field piloting and validation of questions based on statistical criteria. The difficulty of questions is determined a priori by the experience of question designers and the Computerized Testing Committee, resulting in a lack of data on candidates' response times.

- The randomization of test questions, varying among candidates, may lead to discrimination, as it could label challenging questions as first questions for some of the candidates.

Conclusions

In this comprehensive review, we emphasize on the recognition of the need for efficiency, standardization, and accuracy in evaluating student achievements we advocate for the incorporation of an e-assessment system in the State Exam with a comprised balanced mixture of Multiple-Choice Questions (MCQs) and Short Answer Questions (SAQs). The automated grading system ensures swift and objective assessment, crucial for maintaining reliability, particularly in scenarios where multiple markers are involved. This approach not only streamlines the grading process but also contributes to a more consistent and fair evaluation across the extensive student cohort.

Transitioning from paper-based assessments to e-assessments emerges as a prudent move for several compelling reasons. The attainment of a sample size exceeding the minimum requirement alleviates concerns regarding poor psychometric analysis, ensuring robust and reliable evaluations. The flexibility to employ either raw scores or Item Response Theory (IRT) for student proficiency assessments, coupled with the implementation of accuracy tests, enhances the precision of evaluation methods. Moreover, the shift to e-assessment provides exam developers with detailed information, facilitating ongoing improvements in exam reliability. Psychometricians benefit from enhanced test analysis and statistical reliance, contributing to a more comprehensive understanding of assessment outcomes. This transition fosters increased student confidence, as they engage with a more transparent and efficient evaluation process.

Setting rigorous standards for markers is paramount in ensuring the validity and reliability of any assessment process. The act of grading, inherently subjective, requires a structured framework to mitigate the potential for inconsistency and bias among different markers. Standardized marking criteria serve as a benchmark, aligning the assessment with predefined learning objectives and expectations. This not only guides markers in their evaluations but also provides a transparent and accountable system for both educators and students. Efficient calibration of raters with a global rating scale is paramount to ensure the reliability and fairness of assessments. Rigorous training sessions, including clear guidelines and illustrative examples, form the foundation for establishing a shared understanding among raters. Calibration exercises, where raters independently evaluate the same samples, provide an opportunity to address discrepancies and enhance alignment. Despite some attempts, the absence of clear investments and established policies poses a challenge in achieving an efficient and standardized approach in this regard.

Overall, adopting these recommendations not only addresses existing limitations but also represents a strategic advancement toward more effective, reliable, and student-centric evaluation practices.

References

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. Journal of the Medical Library Association: JMLA, 103(3), 152–153. doi: 10.3163/1536-5050.103.3.010

Ahrens, A., & Zaščerinska, J. (2014). A framework for selecting sample size in educational research on e-business application. Proceedings of the 11th International Conference on E-Business. Presented at the International Conference on e-Business, Vienna, Austria. DOI: 10.5220/0005020000390046

Al-Rukban, M. O. (2006). Guidelines for the construction of multiple-choice questions tests. Journal of Family and Community Medicine, 13(3), 125-133. PMID: 23012132; PMCID: PMC3410060. https://pubmed.ncbi.nlm.nih.gov/23012132/

Barker, C., Pistrang, N., & Elliott, R. (2015). Research Methods in Clinical Psychology: An Introduction for Students and Practitioners (3rd ed.). Oxford, UK: John Wiley & Sons, Ltd.

Besekar, S., Jogdand, S., & Naqvi, W. (2023). Sample size in educational research: A rapid synthesis. F1000Research, 12, 1291. https://doi.org/10.12688/f1000research.141173.1

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain (pp. 201-207). New York: McKay.

Borg, W. R. (1987). Applying educational research: A practical guide for teachers. New York: Longman. https://searchworks.stanford.edu/view/1238399

Borg, W. R., & Gall, M. D. (1989). Educational research. An introduction (5th ed.). White Plains, NY: Longman.

Boruch, R. F. (2003). Randomized field trials in education. In T. Kellaghan & D. L. Stufflebeam (Eds.), International Handbook of Educational Evaluation (Vol. 9, pp. 135-152). Springer. https://doi.org/10.1007/978-94-010-0309-4_9

Center for Educational Services (CES), Regulated Profession Item Fund

http://qsha.gov.al/DPSH/fondi_pyetjeve.html

Center for Educational Services (CES), State Matura item Fund

http://qsha.gov.al/DPSH/fondet_html/Matura%20Mjekesi/Fondi%20i%20Pyetjeve%20i%20Provimit%20te%20Informatizuar%20te%20Mjekesise.html

Coolican, H. (2014). Research methods and statistics in psychology (6th ed.). https://doi.org/10.4324/9780203769836

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. Dental Press Journal of Orthodontics, 19(4), 27–29. doi: 10.1590/2176-9451.19.4.027-029.ebo

Hackshaw, A. (2008). Small studies: strengths and limitations. The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology, 32(5), 1141–1143. DOI: 10.1183/09031936.00136408

Hijji, B. M. (2017). Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study. The Journal of Nursing Education, 56(8), 490–496. https://doi.org/10.3928/01484834-20170712-08

Holland, J., O'Sullivan, R., & Arnett, R. (2015). Is a picture worth a thousand words: an analysis of the difficulty and discrimination parameters of illustrated vs. text-alone vignettes in histology multiple choice questions. BMC Medical Education, 15(1), 184. DOI: 10.1186/s12909-015-0452-9

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. Psychology (Irvine, Calif.), 09(08), 2207–2230. DOI: 10.4236/psych.2018.98126

Mee, J., Pandian, R., Wolczynski, J., Morales, A., Paniagua, M., Harik, P.,Clauser, B. E. (2023). An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Advances in Health Sciences Education: Theory and Practice. https://doi.org/10.1007/s10459-023-10266-3

MoES Regulated Professions in Albania Law 2009,

https://arsimi.gov.al/wp-content/uploads/2017/10/LIGJ_NR_10_171_PRR.pdf

MoES, 2023, https://arsimi.gov.al/wp-content/uploads/2023/12/Urdher-i-perbashket-MFE-nr.661-date-01.12.2023-Rregullore-e-MSH.pdf Article 10, point 3

Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: ten main limitations and recommendations to improve future research practices. Psicologia, 30(1), 3. https://doi.org/10.1186/s41155-016-0057-1

Nguyentan, D.-C., Gruenberg, K., & Shin, J. (2022). Should multiple-choice questions get the SAQ? Development of a short-answer question writing rubric. Currents in Pharmacy Teaching & Learning, 14(5), 591–596. https://doi.org/10.1016/j.cptl.2022.04.004

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

PISA 2025 technical standards

https://www.oecd.org/pisa/pisaproducts/PISA_2025_Technical_Standards.pdf

Przymuszała, P., Piotrowska, K., Lipski, D., Marciniak, R., & Cerbin-Koczorowska, M. (2020). Guidelines on Writing Multiple Choice Questions: A Well-Received and Effective Faculty Development Intervention. SAGE Open, 10(3).

https://doi.org/10.1177/2158244020947432

Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. Advances in Health Sciences Education: Theory and Practice, 24(1), 141-150. https://doi.org/10.1007/s10459-018-9855-9

Sam, A. H., Westacott, R., Gurnell, M., Wilson, R., Meeran, K., & Brown, C. (2019). Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. BMJ Open, 9(9), e032550. DOI: 10.1136/bmjopen-2019-032550

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), 116–131.

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/

Slekar, T. D. (2005). Without 1, Where Would We Begin? Small Sample Research in Educational Settings. Journal of Thought, 40(1), 79–86. http://www.jstor.org/stable/42589814

State Matura regulation, 2022 https://arsimi.gov.al/wp-content/uploads/2023/01/Udhezim-Nr.32-date-23.12.2022-Matura-shteterore-pdf.pdf

State Matura regulation, 2024

https://arsimiparauniversitar.gov.al/wp-content/uploads/2024/01/Urdher-i-perbashket-MFE-nr.661-date-01.12.2023-Rregullore-e-MSH-njoftimi-1.pdf

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Medical Research Methodology, 18(1), 148.

https://doi.org/10.1186/s12874-018-0594-7

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press