

Assessment and evaluation through e-tests: experiences and reliability in higher education within the scope of probability theory course

Gabriella Papp^{a,b}

^aUniversity of Debrecen, Institute of Mathematics,
Doctoral School of Mathematical and Computational Sciences
p.gabica.17@gmail.com

^bFerenc Rakoczi II Transcarpathian Hungarian University
papp.gabriella@kmf.org.ua

Abstract. Crises accelerated the digital shift in higher education, making e-testing essential at Ferenc Rakoczi II Transcarpathian Hungarian University. While e-testing in mathematics is limited by its focus on final results over logic, the demand for rapid, objective feedback and the reduction of administrative burdens make computer-based systems essential in modern higher education.

The present research is an exploratory case study focusing on the field of combinatorics, which serves as the foundation of probability theory and requires a high level of abstraction. The subjects of the study are students majoring in Accounting and Taxation at Ferenc Rakoczi II Transcarpathian Hungarian University, for whom a stochastic mindset is essential for their professional future. This study analyzes not only immediate performance but also the reliability of the assessment through the method of delayed testing. This approach enables the examination of “retained knowledge”, highlighting the extent to which knowledge acquired in a digital environment is integrated into long-term memory and can be considered genuine, applicable competence within an educational environment affected by crisis situations.

Keywords: mathematics education, higher mathematics, probability theory, e-testing, test reliability

2020 Mathematics Subject Classification: 97D60, 97U50, 97K20, 97K50, 97C30

1. Introduction

Digital transformation has reached every segment of higher education; however, integrating computer-based assessment and e-testing into mathematics education presents unique methodological challenges for institutions. Today, the flexibility and sustainability of educational systems are inseparable from digital solutions. In this environment, scientific research in this field has become essential, as digital alternatives require preparation distinct from traditional measurements to ensure that mathematical knowledge is assessed according to established standards. Fundamental questions regarding assessment elicit various responses worldwide. The most critical primary questions are: What is the purpose of the evaluation? Why (do we evaluate)? These are followed by several practical questions: Who? Where? When? How? And with what tools (does one evaluate)? [13, 15].

The relevance of this pilot research is further underscored by the methodological dilemma that, due to the abstract nature of mathematics – and specifically probability theory – many question the validity of electronic testing. The findings of this study create direct practical value: on one hand, they provide educators with a framework and empirical foundation to overcome technical constraints (such as the difficulties of online examination or random task generation); on the other hand, they provide evidence that, following proper design principles, genuine stochastic knowledge and “retained knowledge” can be measured on online interfaces.

The primary aim of this research is to investigate a specialized computer-based measurement and evaluation framework built upon the principles of classical and modern test theories (such as reliability and validity), providing a methodological response to the digital challenges of higher mathematics education. A central element of the study is the identification of adaptability mechanisms for custom-developed e-tests, with a particular focus on how abstract mathematical concepts – specifically within the field of probability theory – can be effectively transitioned into a digital environment.

The research aims to uncover the systemic regularities that determine the effectiveness and scientific credibility of electronic assessment. This includes analyzing the extent to which technological solutions applied during the e-testing process can support students’ long-term knowledge retention, namely the consolidation of “retained knowledge”. The resulting model of this research enables the methodological renewal of probability theory instruction, ensuring the validity of digital measurements within the requirements of the higher education system. To achieve the aim, the following objectives were set:

- To define task types tailored to the specific cognitive requirements of combinatorics, taking into account the various levels of mathematical abstraction;
- To pilot the assessment tool among the target group while ensuring controlled measurement conditions;
- To examine whether the test is suitable for both embedded formative feedback and the objective recording of final summative performance;

- To re-evaluate student performance after a specified period to identify the extent of “retained knowledge” and the degree of long-term knowledge retention;
- To calculate internal consistency indicators (e.g., Cronbach’s alpha) based on the collected data in order to minimize measurement error.

2. Literature review

In the process of pedagogical assessment, a student’s observed behavior and products are compared to a scale based on predetermined criteria and reference points using a measurement tool to quantify the results achieved [19]. One of the most common forms and specific areas of pedagogical assessment is educational attainment testing (knowledge-level measurement). The theory of attainment testing evolved from the psychological theory of testing. Knowledge-level tests measure psychological traits and a specific form of knowledge – typically the knowledge involved in formal school education [6].

Based on the definition in the Lexicon of Pedagogy, [28] defines assessment as a process aimed at the systematic examination or judgment of the value of a phenomenon, event, or object. During a background research, readers can encounter various formulations of the concept of a “test”:

- A test (originally meaning “trial”) is a typical measurement tool in pedagogy. In pedagogical and psychological terminology, it refers to the assessment of a specific trait through testing [7].
- A test is a term used to describe standardized procedures through which certain behaviors can be described or measured to assign categories or scores to individuals or events in a predetermined manner [16].
- A test is a measurement instrument that assesses the psychological phenomenon under study on an appropriate scale [28].
- Knowledge-level tests examine and measure what a student is capable of when their total knowledge is aggregated [5].
- A didactic test is a type of examination in which students typically respond in a specified manner (closed-ended items with structured responses) and, in fewer cases, construct their own answers (open-ended items) [10].

For our purposes, the common emphasis across these definitions is measurement, whereby students’ performance and abilities become measurable, comparable, and quantifiable.

2.1. Key elements of test construction

In a narrower sense, pedagogical assessment is a method of teaching and learning [15]. Assessment is no longer merely a conclusion but rather the theory and practice of organized and differentiated feedback of pedagogical information [24, 25], which is built upon the relationship between the objective and the achieved impact, as well as the input and evaluation of various levels and elements of teaching and learning [15]. Pedagogical assessment, when applied as a teaching method, provides information regarding the level of achievement of set goals [28]. This enables the evaluation of not only student activities but also the entire teaching-learning process and its effectiveness, including every factor involved in this process. Assessment methods applied within electronic learning environments provide opportunities for immediate feedback [24], which is of critical importance because, in the absence of feedback, erroneous information may become consolidated [26].

Within digital assessment environments, the test is one of the most widespread and dynamically evolving tools for knowledge-level measurement. Objectivity of the assessment is ensured by precise and equal scoring rules and by scoring (marking) identical answers in the same way, so that its results are independent of who does the correcting, coding, i.e. scoring of the results. The objectivity of a test is achieved by the correct selection of items, the clear formulation of questions and response options, adequate coverage of the material, and the elimination of all doubt regarding the correct answers to the items. This objectivity ensures that the test results remain independent of the person administering the assessment. The objectivity of the testing process is achieved by providing equal testing conditions for all students being tested [6, 10, 25].

One of the most important properties of a test is its reliability [6] (Csapó, 2004). Test reliability is an indicator of the accuracy and reliability of measurement [10], which estimates measurement errors at the group level. For a test with high reliability, measurement errors are low. Reliability can be estimated using different methods [19]. Cronbach created the formula for calculating the universal index in 1951:

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\sum_i V_i}{V_t} \right) \quad (2.1)$$

where n is the number of items, V_t is the variance of the total scores, and V_i is the variance of item scores [4, 23].

Generally speaking, the greater the number of items, the more reliable the measurement [19]. If the number of items is small or the average correlation is low, Cronbach's alpha will also be low [16].

The more tasks and curriculum aspects are examined, the more reliable the test results become. Consequently, formal tests typically consist of a relatively large number of items. From a practical standpoint, a test should comprise at least 20–30 items, and increasing the number of items can further enhance reliability [10, 11, 19].

Reliability can be quantified by a value between 0 and 1 [8]. Ideally, the rela-

bility of a test is 1, which implies that if such a test were administered repeatedly, every student would achieve the same result for each item as they did during the first measurement. In practice, however, a reliability of 1 does not occur [10]. CORTINA [3] demonstrated through various hypothetical examples that alpha can remain high (greater than 0.7) despite low item-intercorrelations and multidimensionality, provided that a sufficient number of items are available [3, 23]. Although the standards for establishing a “good” α coefficient are entirely arbitrary, many methodologists suggest a minimum α coefficient between 0.65 and 0.85 (or in many cases higher); α coefficients lower than 0.5 are generally unacceptable [10, 11, 14].

A key question is whether the test measures what it was intended to measure. PARÁZSÓ [25] emphasizes that without the validity of the measurement, the results cannot be used for pedagogical decision-making. A good test has a differentiating effect; that is, it distinguishes between lower-performing and higher-performing students. The validity of knowledge-level tests can be ensured by applying a test-construction technique that allows for the precise mapping of the curriculum onto the instrument used for measuring and testing said material [6].

CSAPÓ [6] refers to the relationship between task reliability and validity. According to DEMKANIN ET AL. [10], reliability is a necessary but not sufficient condition for validity. For a test to be valid, it must also be reliable [8]; thus, it must possess a high reliability index. This relationship does not hold in reverse: if a test’s validity is weak, its reliability can still be very high [6]. Reliability and accuracy are therefore prerequisites for validity, but they are not the only prerequisites [10].

During the design process, it is beneficial to apply the dimensions of knowledge and cognitive processes in a varied manner when creating items [27]. Currently, the taxonomies of Niemierko and Bloom are the most widely recognized frameworks applied for this purpose. Each taxonomy is in fact a hierarchically arranged classification [10].

Bloom’s Taxonomy is a classification of various objectives and skills that educators set for their students (learning objectives). The system was proposed in 1956 by Benjamin Bloom, an educational psychologist at the University of Chicago [27]. Compared with Niemierko’s taxonomy, Bloom’s taxonomy provides a breakdown of higher-order cognitive processes and is more detailed regarding subcategories. In contrast, Niemierko’s taxonomy aggregates all higher-level logical-cognitive processes into a single category (non-specific transfer) [10]. Table 1 shows the two-dimensional table of Bloom’s Taxonomy.

The content validity and structural design of the assessment tool are based on international and domestic literature on pedagogical evaluation. A detailed scientific analysis of the two-dimensional matrix and the cognitive levels used for item classification is provided in the taxonomic revision by ANDERSON ET AL. [2], as well as in the research by DEMKANIN ET AL. [10] and SHABATURA [27]. Designing an ideal assessment tool requires extensive coverage of the two-dimensional taxonomic matrix. This approach – which combines the dimension of cognitive processes (e.g., remember, understand, apply) with types of knowledge (e.g., fac-

Table 1. Summary table of the Bloom’s taxonomy [2, 10, 27].

			The cognitive process dimension					
			1	2	3	4	5	6
			Remember	Understand	Apply	Analyze	Evaluate	Create
The knowledge dimension	A	Factual knowledge						
	B	Conceptual knowledge						
	C	Procedural knowledge						
	D	Metacognitive knowledge						

tual, conceptual, procedural) – enables the precise and differentiated measurement of student knowledge and specific competencies. A diverse classification ensures that the test evaluates not only basic recall but also the complex analytical skills essential for higher-level mathematical abstraction.

2.2. The relevance of the topic

Within the theoretical framework of pedagogical measurement and evaluation, literature fundamentally distinguishes between three main modes of assessment: diagnostic, formative, and summative [13].

The academic discourse regarding digital assessment in mathematics reveals a nuanced tension between technological benefits and pedagogical limitations. On one hand, JAKAB [17] and DEBRENTI [9] emphasize the necessity of physical presence and analog tools, arguing that the abstract nature of complex mathematical problems is often better grasped through direct, hands-on interaction. Their findings suggest that digital interfaces may occasionally hinder the depth of conceptual understanding. According to PASHLER ET AL. [26] regular testing halts forgetting because students utilize quizzes for a deeper understanding of content and for the consolidation of knowledge [12, 26].

Conversely, MUZSNAY ET AL. [20] highlight the long-term cognitive advantages of digital platforms, demonstrating that regular e-testing via retrieval practice effectively mitigates the “forgetting curve” and ensures knowledge retention. In their experiment, measured students’ knowledge three months after the conclusion of the course, confirming long-term retention [20]. This resonates with Pashler et al principle, which states that the temporal distribution of learning and testing sessions (spacing effect) increases memory durability [26].

This potential for continuous engagement is further supported by NÉGYESI [21], who points out that adaptive e-learning environments allow for highly personalized formative assessment, where the system dynamically adjusts to the student’s needs. Worked examples and quizzes reduce cognitive load [1], allowing students to use

quizzes to plan their further learning [12].

However, NÉGYESI [21] also introduces a critical caveat regarding the “blurring boundaries” between assessment types. He warns that learning data and adaptive quizzes, originally intended for developmental (formative) purposes, are frequently repurposed for summative evaluation. This shift aligns with USISKIN [29] broader critique, suggesting that an overreliance on automated testing may distort instructional practices and reduce mathematics to a series of measurable, mechanical units. Ultimately, the literature suggests that while e-testing may struggle to capture the initial process of discovery, it is an unparalleled tool for supporting long-term retention and providing the objective feedback necessary for modern, resilient higher education – provided that the delicate balance between formative support and summative pressure is maintained.

According to KORENOVA ET AL. [18] future primary school teachers must develop new teaching approaches in a rapidly changing digital world. This necessity is also reflected in the analysis by JAKAB [17], which notes that teachers underwent a forced but intensive digital development due to the impact of various crises. KORENOVA ET AL. [18] and DEBRENTI [9] maintain that technology opens new dimensions in education [9, 17, 18].

3. Materials and methods

The practical piloting of the custom-developed electronic assessment tool and the associated data collection were conducted among the students of the Ferenc Rakoczi II Transcarpathian Hungarian University. The methodological framework of this research was established through systematic and objective measurement procedures. The target group of the study consisted of a small group ($n=10-11$) of students from the Accounting and Taxation major, representing a population in economic education that requires a foundation in mathematics and statistics.

The test items were predominantly scored dichotomously, awarding 1 point for a correct response and 0 points for an incorrect one. Partial credit was applied exclusively to matching and multiple-response items. This scoring scheme facilitated the calculation of classical test theory metrics, specifically item difficulty (p) based on classical probability, and the item discrimination index (D) determined via the upper-lower group method.

The e-tests used in the assessment are not published in the appendix. On the one hand, this is to maintain the security of the evaluation system – as the tests are accessible to students exclusively during the official testing period – and on the other hand, due to the nature of the item banks, individual items do not contain direct references to the source problem sets or the modifications made to them.

3.1. Item design of the e-tests

In this phase of the pilot research, we examined the diagnostic and differentiating power of the custom-developed electronic assessment tools. Although it was not

possible to administer a pre-test on the topic of combinatorics, a topic-closing (1st test) and a delayed test (2nd test) made it possible to measure the knowledge that students had actually retained, while ensuring the reliability of the e-test.

Table 2. Structural summary of test items.

№	Topic			Bloom taxonomies	Items						Difficulty		
	Permutation	Variation	Combination		Multiple Choice (Single)	Multiple Choice (Multiple)	True-False	Matching	Short Answer	Fill-in-the-blank	Easy	Medium	Difficult
1	+			B1						+	+		
2			+	C4			+				+		
3	+	+	+	B1				+			+		
4	+			B3					+		+		
5	+			C3	+							+	
6	+			B2					+			+	
7	+			C3	+								+
8		+		C3					+			+	
9		+		C3	+							+	
10		+		C2		+						+	
11		+		C4					+		+		
12			+	C3	+							+	
13			+	B3		+					+		
14			+	C3	+								+
15			+	C3	+								+
16			+	C4			+				+		
17			+	C3					+			+	
18			+	B1						+	+		
19			+	C4	+							+	
20			+	D3					+				+
T	6	5	11	20	7	2	2	1	6	2	8	8	4

The curriculum comprised the three fundamental elements of combinatorics – per-mutation, variation, and combination – including their repetitions, which were addressed during practical sessions through collaborative brainstorming and individual problem-solving.

To ensure content validity and a differentiated examination of cognitive abilities,

Table 3. Structural summary of the revised e-test.

No.	Rearrangement	Topic			Bloom taxonomies	Items						Difficulty		
		Permutation	Variation	Combination		Multiple Choice (Single)	Multiple Choice (Multiple)	True-False	Matching	Short Answer	Fill-in-the-blank	Easy	Medium	Difficult
1	1	+			B1						+	+		
2	2			+	C4			+				+		
3	11		+		C4					+		+		
4	4	+			B3					+		+		
5			+		C3	+						+		
6	13			+	B3		+					+		
7	3	+	+	+	B1				+			+		
8		+			D3	+						+		
9			+		C3		+					+		
10	18			+	B1						+	+		
11	5	+			C3	+							+	
12	6	+			B2					+			+	
13		+			C3					+			+	
14			+		C3	+							+	
15	9		+		C3	+							+	
16	10		+		C2		+						+	
17	19			+	C4	+							+	
18	12			+	C3	+							+	
19	16			+	C4			+					+	
20	17			+	C3					+			+	
21	7	+			C3	+								+
22	8		+		C3					+				+
23	14			+	C3	+								+
24	15			+	C3	+								+
25	20			+	D3					+				+
T		8	8	11	25	10	3	2	1	7	2	10	10	5

the test items were categorized according to Bloom's Taxonomy (see Table 1), with tasks ranging from simple knowledge recall (level A1) to complex analytical and application skills (level D6).

To ensure the credibility and reliability of the results, both research stages were conducted under strictly controlled classroom conditions with continuous supervision to guarantee independent work. The first e-test conducted upon the completion of the combinatorics unit with a sample of 10 students. The tool consisted of a 20-item e-test (see Table 2) with a 45-minute time limit.

The internal arrangement of the test adhered to the principles of systematic progression and continuity. Regarding difficulty scaling, each thematic unit included easy, medium, and difficult tasks as determined by us.

Following a rigorous structural logic, basic-level tasks from all three relevant subject areas were positioned at the beginning of the test in a randomized manner, followed by items of increasing complexity. This strategic sequencing not only optimized the students' cognitive load but also facilitated a precise analysis of the instrument's discriminatory power.

Conducted two months later to analyse the longitudinal effects of e-test supported instruction and the reliability of the self-developed e-test within the group. All 11 students participated in this phase. To prevent rote recall or reliance on a remembered sequence, the second test was expanded to 25 items (see Table 3), and task orders were randomized. Given the expansion, a 60-minute working time was provided.

During the selection and structuring of the new 5 tasks, Bloom's cognitive taxonomy served as the foundation, ensuring content validity and continuity across cognitive levels. The test's internal arrangement adhered to the principles of systematic progression and continuity. Methodologically, basic-level tasks from all three relevant subject areas (e.g., various types of combinatorics) were positioned at the beginning of the test in a randomized manner. These were followed by tasks of medium difficulty and, ultimately, by complex tasks requiring higher-order cognitive operations. This structured sequence optimizes the students' cognitive load while facilitating a more precise analysis of the instrument's discriminatory power.

The item difficulty indices summarized in Tables 2 and 3 are determined by expert judgment based on prior instructional experience.

3.2. Technological framework: The Online Test Pad platform

Aligning with modern higher education methodology, the [Online Test Pad](#) platform was selected as the technological framework for its robust support of digital pedagogy. The platform offers several critical technical and pedagogical advantages:

- **Mathematical precision:** It supports $\text{\LaTeX}$ -based entry, ensuring the professional display of mathematical formulas.
- **Security and integrity:** Features such as randomized task orders, randomized options, disabled clipboards, and strict time limits ensure a controlled assessment environment.
- **Accessibility:** Access via a direct link requires no registration, minimizing measurement errors caused by technical barriers.

- **Objectivity:** The platform allows for the full automation of the assessment process. A uniform scoring system is applied, and grading is performed automatically, guaranteeing standardized results free from instructor subjectivity.

4. Results

4.1. Interpretation and analysis of results

Based on the results of the first e-test, it can be concluded that the tool is suitable for formative feedback and possesses a distinct summative discriminatory effect. The total scores achieved by the students show significant variance, which correlates with student activity and prior knowledge levels observed during the instructional process. The distribution of the results confirms the assumption that the test is capable of objectively isolating different competency levels, in alignment with the findings of Parázso regarding objective measurement and evaluation.

Data were coded and analyzed using Microsoft Excel. The expanded 2nd e-test used to measure retained knowledge holds similar diagnostic value. The analysis of summative scores here also confirms discriminatory power, while the immediate evaluation provided by the platform served as an effective formative tool for the students.

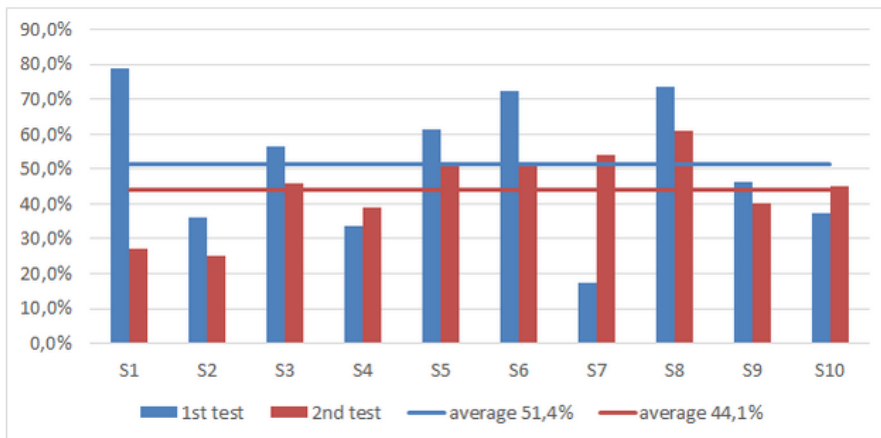


Figure 1. Comparison of results from the first and second (delayed) measurements.

Figure 1 presents a comparative analysis of the results from the two measurement points. The diagram compares the percentage performance of the 10 students (see $S1 \dots S10$) who participated in both stages of the study. This comparison allows for tracking the extent of knowledge retention, as well as the dynamics of knowledge loss or consolidation occurring during the two-month interval. The

individual performance differences shown in the diagram reinforce the test's differentiating power. The students' results are not clustered within a narrow range but cover the entire scale.

The percentage (Figure 1) shifts shown in the diagram represent the loss of knowledge and the retained knowledge occurring over the two-month interval. In alignment with the literature, it can be observed that where students successfully solved items requiring higher-order cognitive operations (Bloom $C3 - D3$) during the first measurement, the ratio of "retained knowledge" experienced during the delayed test also remained higher.

Table 4. Difficulty and discrimination analysis of Test 1 and Test 2.

No.	1st Test	p1	D1	p2	D2	Nature of change
1.	Q1	0.5	0.64	0.73	0.28	Became easier, weaker D
2.	Q2	0.5	0.37	0.36	0.06	Became harder, weak D
3.	Q11	0.1	-0.09	0.09	0.2	Stably hard, improving D
4.	Q4	0.8	0.1	0.91	Zero*	Became too easy
5.	New item	-	-	0.18	0.42	Strong new item
6.	Q13	0.65	0.73	0.68	0.01	Stable with weak D
7.	Q3	0.725	0.54	0.48	0.09	Became significantly harder
8.	New item	-	-	0.55	0.4	Good new item
9.	New item	-	-	0.64	0.11	Average new item
10.	Q18	0.6	0.61	0.5	0.86	Outstanding retention
11.	Q5	0.7	0.61	0.45	0.4	Harder, but good D
12.	Q6	0.7	0.74	0.55	-0.15	Became harder
13.	New item	-	-	0	Zero*	Too hard new item
14.	New item	-	-	0.36	0.6	Strong new item
15.	Q9	0	Zero*	0.09	-0.52	Remains critical
16.	Q10	0.5	0.67	0	Zero*	Surprising decline
17.	Q19	0.4	0.4	0.18	0.57	Harder, but excellent D
18.	Q12	0.6	0.55	0.45	0.2	Became harder
19.	Q16	0.8	0.63	0.55	0.68	Stable and good item
20.	Q17	0	Zero*	0.45	0.24	Surprising improvement
21.	Q7	0.5	0.43	0.18	0.51	Harder, excellent D
22.	Q8	0.5	0.37	0.09	-0.17	Critical decline
23.	Q14	0.4	0.39	0.55	0.47	Easier and better D
24.	Q15	0.5	0.32	0.36	0.04	Harder and weaker D
25.	Q20	0.8	0.67	0.64	0.69	Very stable item

The two-stage measurement conducted during the pilot research highlighted a correlation between the movement of mathematical knowledge and the assessment tool's stability. Based on the results (see Table 4), the item difficulty index ($p1$ and $p2$) showed a general decrease over the two-month interval, serving as direct evidence of the natural forgetting process. While the majority of items fell within

the medium difficulty range during the first measurement ($p \approx 0.5$), several items in the delayed test (such as tasks 10, 17, or 19) became significantly more difficult for the students. This confirms that the long-term consolidation of abstract combinatorial concepts remains uncertain without regular reinforcement.

The evolution of the discrimination index (Table 4, $D1$ and $D2$) presents an even more complex picture. During the delayed measurement, several items lost their ability to separate high-performing and low-performing students effectively; due to increasing uncertainty, students were more frequently forced to guess, introducing statistical “noise” into the system. However, certain key items (such as tasks 10, 19, and 25) maintained their high discriminatory power, proving that structured test design based on Bloom’s Taxonomy is capable of identifying genuine, deeply rooted knowledge even under deteriorating measurement conditions.

4.2. Critical analysis of reliability and validation of measurement results

Data were coded and analyzed using Microsoft Excel, where the internal consistency and reliability of the assessment tool were quantified through Cronbach’s alpha. The coefficient was derived via a two-way ANOVA without replication, utilizing the “Data Analysis” add-in to facilitate the calculation of formula (2.1). This approach utilizes the mean square of rows (MS_{Rows} , representing variance between subjects) and the mean square of error (MS_{Error} , representing residual variance) extracted directly from the ANOVA tables.

In the measurement conducted immediately after the end of the instructional process, the Cronbach’s alpha value was $\alpha = 0.80$. Based on the standards of the literature, this value indicates good/acceptable reliability (see in subsection 2.1). This result scientifically proves that:

- The internal consistency of the custom-developed, 20-item e-test is appropriate;
- The task set designed on the Online Test Pad platform consistently measured the students’ competencies in combinatorics;
- The assessment tool was suitable for summative differentiation among students.

In contrast, during the expanded measurement conducted two months later, the coefficient was significantly lower, at $\alpha = 0.46$. Although this value falls below the 0.70 threshold typically expected in research, it can be explained by the following scientific factors within the present experimental context:

- Cronbach’s alpha is extremely sensitive to small sample sizes.
- As knowledge faded, students were more frequently forced to estimate or guess, which introduces statistical “noise” (error variance) into the system, directly reducing the alpha value.

The discriminatory power of the assessment tools is illustrated by Figure 1, which presents the individual student performance at the two measurement points, and Table 4, which details the difficulty indices and discrimination levels of the items. Based on the data, it can be concluded that the students' results were dispersed across a wide interval in both tests. From a scientific perspective, this significant variance proves that the tests successfully distinguished between students with different competency levels. The distribution of the scores achieved supports the differentiating power of the items, enabling the objective ranking of student knowledge levels within the field of combinatorics.

5. Discussion

The obtained results can primarily be explained by the synergy between cognitive psychology and digital pedagogy. The high reliability observed during the first measurement ($\alpha = 0.80$) confirms that the structured e-test, based on Bloom's Taxonomy, is capable of providing an objective reflection of immediate knowledge levels. Two months later, a decline in reliability becomes apparent ($\alpha = 0.46$), so the reliability of the e-test cannot be statistically substantiated. The decline in statistics recorded is attributable to natural forgetting processes, the statistical sensitivity of small-sample measurements, and individual variance in knowledge retenti

The Online Test Pad-based [22] assessment tool used in this research, equipped with LaTeX support, offers numerous advantages over traditional alternatives. Most importantly, it eliminates instructor subjectivity, provides immediate feedback, and minimizes the possibility of copying through its parameterization features (randomized order, shuffling).

The results of the pilot research – although statistically limited due to the small sample size – provide practical evidence that the adaptation of digital measurement and evaluation to this group in this field of mathematics is still in its infancy. Thus, the research merely presents a method, which validates the existence of the gaps identified in the literature: the tension between the abstraction of probability theory and the statistical stability of e-tests remains one of the greatest methodological challenges in the field.

When applying the proposed method in practice, the following limitations must be considered:

- the validity of the results is only guaranteed in a controlled, supervised environment (with personal presence);
- in the case of remote completion, external aids may distort the data;
- platform stability and a continuous internet connection are prerequisites for an uninterrupted measurement process.

The primary limitation of the case study is the sample size, which restricts statistical generalizability and contributed to the low Cronbach's alpha value in

the delayed measurement. Furthermore, the pilot research did not address the individual influencing effects of different learning styles or the degree of digital anxiety, a deeper analysis of which would provide a more accurate picture of the background of student performance.

The research could be further developed by incorporating intelligent algorithms that dynamically change the difficulty of the next task based on the student's responses, as well as by involving multiple institutions and majors to increase statistical validity.

6. Conclusion

The case study examines how a structured e-test behaves in one university cohort and what patterns appear in delayed testing. During the pilot research, we successfully developed a digital assessment tool based on the Online Test Pad platform, which, thanks to LaTeX-based mathematical character handling and Bloom's Taxonomy, is suitable for measuring in this group's competencies in probability theory. A distinguishing feature of the developed system is the deliberate task hierarchy according to cognitive levels. This allowed for the systematic distribution of easy, medium, and difficult tasks throughout the measurement process, ensuring that the assessment examines not only reproductive knowledge but also application skills.

The primary characteristic of the results is the correlation between the two measurement points (the immediate post-test and the delayed test conducted 2 months later). The measurement confirmed that the e-test is suitable for the differentiated evaluation of students: the results obtained during the first measurement faithfully reflected the students' classroom performance and preparation. Immediate feedback and objective, instructor-independent grading ensured the bias-free recording of performance.

Quantitative analysis using ANOVA revealed the temporal dynamics of the assessment tool's reliability: the excellent internal consistency observed during the first measurement ($\alpha = 0.80$) decreased significantly for the delayed measurement ($\alpha = 0.46$), which is closely linked to the weakening of the items' differentiating effect. While the success of the first stage is supported by the strictly structured item design based on Bloom's Taxonomy, the subsequent statistical decline in reliability and discrimination explained is attributable to the divergent nature of natural forgetting processes and the increased guessing factor resulting from knowledge loss.

As there is limited research available concerning higher education mathematics, it remains an unanswered question how electronic testing can be adapted in higher education mathematics subjects so that it maintains its statistical stability even during delayed measurements. Despite the lower alpha value, the content validity and discriminatory power of the test continue to provide valuable feedback in practical education regarding the profile of retained knowledge.

References

- [1] S. M. ADENJII, P. BAKER: *Effects of worked example on students' learning outcomes in complex algebraic problems*, International Journal of Instruction 16.2 (2023), pp. 229–246, DOI: [10.29333/iji.2023.16214a](https://doi.org/10.29333/iji.2023.16214a).
- [2] L. W. ANDERSON, D. R. KRATHWOHL, ET AL.: *A revision of Bloom's taxonomy of educational objectives*, United States: Longman, 2001, URL: https://www.researchgate.net/publication/235465787_A_Taxonomy_for_Learning_Teaching_and_Assessing_A_Revision_of_Bloom's_Taxonomy_of_Educational_Objectives.
- [3] J. M. CORTINA: *What is coefficient alpha? An examination of theory and applications*, Journal of Applied Psychology 78.1 (1993), pp. 98–104, DOI: [10.1037/0021-9010.78.1.98](https://doi.org/10.1037/0021-9010.78.1.98).
- [4] L. J. CRONBACH: *Coefficient alpha and the internal structure of tests*, Psychometrika 16.3 (1951), pp. 297–334, DOI: [10.1007/BF02310555](https://doi.org/10.1007/BF02310555).
- [5] B. CSAPÓ: *A tanulói teljesítmények értékelésének mérési módszerei*, Budapest, Veszprém: Művelődési Minisztérium Vezetőképző és Továbbképző Intézete; Veszprém Megyei Pedagógiai Intézet, 1988.
- [6] B. CSAPÓ: *Tudásszintmérő tesztek*, in: Falus Iván (Szerk.) Bevezetés a pedagógiai kutatás módszereibe, Budapest: Műszaki K., 2004, pp. 277–316, URL: <https://publicatio.bibl.u-szeged.hu/11830/>.
- [7] B. CSAPÓ, G. MOLNÁR, K. R. TÓTH: *A papíralapú tesztektől a számítógépes adaptív tesztelésig*, Iskolakultúra 3-4 (2008), pp. 3–16, URL: <http://real.mtak.hu/id/eprint/5635>.
- [8] C. CSÍKOS, M. B. NÉMET: *A tesztekkel mérhető tudás*, in: Csapó Benő (szerk.) Az iskolai tudás, Budapest: Osiris Kiadó, 2002, pp. 91–123, URL: https://publicatio.bibl.u-szeged.hu/11931/1/CsBeno_Iskolai_tudas_2002.pdf.
- [9] E. DEBRENTI: *Using digital game-based learning in mathematics education: A case study with teacher training students*, International Journal for Technology in Mathematics Education 31.3 (2024), pp. 153–162, DOI: [10.1564/tme_v31.3.06](https://doi.org/10.1564/tme_v31.3.06).
- [10] P. DEMKANIN, M. HAJDÚK, H. HANULJAKOVA, T. KUBIŠ, M. LAPITKA, M. MALČÍK: *Metodika tvorby testových úloh a testov*, Bratislava: NÚCEM, 2015, URL: https://www.researchgate.net/publication/349279914_Metodika_tvorby_testovych_ulo_h_a_testov.
- [11] R. L. EBEL, D. A. FRISBIE: *Essentials educational measurement*, New Delhi: Prentice Hall, 1991, URL: https://ebookppsunp.wordpress.com/wp-content/uploads/2016/06/robert_1-ebel_david_a-frisbie_essentials_of_edbookfi-org.pdf.
- [12] F. FEUDEL, A. UNGER: *Students' Strategic Usage of Formative Quizzes in an Undergraduate Course in Abstract Algebra*, International Journal of Research in Undergraduate Mathematics Education 10 (2024), pp. 351–379, DOI: [10.1007/s40753-022-00194-9](https://doi.org/10.1007/s40753-022-00194-9).
- [13] D. K. M. FÜZNÉ: *A természetismeret tanításának módszertana*, Szeged: Szegedi Tudományegyetem; Juhász Gyula Pedagógusképző Kar; Alkalmazott Pedagógiai Intézet; Tanítóképző Tanszék, 2016, URL: <https://eta.bibl.u-szeged.hu/2204/2/index.html>.
- [14] C. GOFORTH: *Using and Interpreting Cronbach's Alpha*, University of Virginia Library (2015), URL: <https://library.virginia.edu/data/articles/using-and-interpreting-cronbachs-alpha>.
- [15] M. HERCZ: *A pedagógiai értékelés gyakorlata*, in: Bábosik István és Torgyik Judit (szerk.): Pedagógusmesterség az Európai Unióban, Budapest: Eötvös Kiadó, 2007, pp. 191–215, URL: https://www.researchgate.net/publication/314225676_A_pedagogiai_ertekeles_gyakorlata.
- [16] I. HIDEKGUTI, K. BALÁZS: *Tesztelmélet*, in: Balázs Katalin, Kovács Judit, Münnich Ákos (Szerk.): Pszichológiai Módszertani Tanulmányok, Debrecen: Debreceni Egyetemi Kiadó, 2015, pp. 65–95, URL: https://www.researchgate.net/publication/305678594_Tesztelmélet.

- [17] E. JAKAB: *Two crises, one digital transformation: Digital practices of Hungarian mathematics teachers in Transcarpathia during remote teaching*, Bulletin of Science and Education 11.41 (2025), pp. 1476–1496, DOI: [10.52058/2786-6165-2025-11\(41\)-1476-1496](https://doi.org/10.52058/2786-6165-2025-11(41)-1476-1496).
- [18] L. KORENOVA, M. KOŽUCHOVÁ, J. DOSTÁL, Z. LAVICZA: *Applications of Augmented Reality Apps in Teaching Technical Skills Courses*, in: *Augmented Reality in Educational Settings*, Brill, 2020, pp. 383–409, DOI: [10.1163/9789004408845_017](https://doi.org/10.1163/9789004408845_017).
- [19] E. K. MOLNÁR, T. VÍGH: *A tantervelmélet és a pedagógiai értékelés alapjai*, „Mentor(h)áló 2.0 Program” TÁMOP-4.1.2.B.2-13/1-2013-0008 projekt, 2013, URL: http://www.jgypk.hu/mentor_halo/tananyag/Tantervelmlet_s_a_pedagogiai_rtkels_alapjai/.
- [20] A. MUZSNAY, C. SZABÓ, J. SZEIBERT: *Retrieval practice – a tool to be able to retain higher mathematics even 3 months after the exam*, Annales Mathematicae et Informaticae (2024), pp. 218–227, DOI: [10.33039/ami.2024.05.002](https://doi.org/10.33039/ami.2024.05.002).
- [21] P. NÉGYESI: *Adaptivitást támogató elektronikus tanulási környezet fogalmi keretei a számelmélet speciális témaköreinek tanításában*, in: *Kihívások és megoldások a XXI. század pedagógiájában*, Eger: Eszterházy Károly Katolikus Egyetem Líceum Kiadó, 2021, pp. 89–101, URL: <https://publikacio.uni-eszterhazy.hu/7158/>.
- [22] ONLINETESTPAD, URL: <https://onlinetestpad.com/>.
- [23] P. PANAYIDES: *Coefficient Alpha: Interpret With Caution*, Europe’s Journal of Psychology 9.4 (2013), pp. 687–696, DOI: [10.5964/ejop.v9i4.653](https://doi.org/10.5964/ejop.v9i4.653).
- [24] L. T. PARÁZSÓ: *Online teszt és tudásszintmérés*, in: *Agria Media 2011 : „Az információk társadalom az alkotó tudás társadalmának digitális előtere”*, Eger: Eszterházy Károly Főiskola Líceum Kiadó, 2012, pp. 117–125, URL: <http://publikacio.uni-eszterhazy.hu/id/eprint/5952>.
- [25] L. T. PARÁZSÓ: *Teljesítményértékelés elektronikus tanulási környezetben*, in: *A kultúraváltás hatása az egyéni kompetenciákra: a digitális kompetencia modelljei*, Eger: Eszterházy Károly Főiskola Líceum Kiadó, 2020, pp. 119–142, URL: <http://publikacio.uni-eszterhazy.hu/6815/>.
- [26] H. PASHLER, D. ROHRER, N. J. CEPEDA, S. K. CARPENTE: *Enhancing learning and retarding forgetting: Choices and consequences*, Psychonomic Bulletin & Review 14 (2007), pp. 187–193, DOI: [10.3758/BF03194050](https://doi.org/10.3758/BF03194050).
- [27] J. SHABATURA: *Using Bloom’s Taxonomy to Write Effective Learning Objectives*, University of Arkansas, Assignments & Measuring Student Learning (2022), URL: <https://tips.uark.edu/using-blooms-taxonomy/>.
- [28] J. SZÉKELY: *Mérés-értékelés a pedagógiában oktatási segédanyag integrációs szakmai feladatokra és pedagógus-szakvizsgára felkészítő szakirányú továbbképzési szak programhoz*, Eger: Eszterházy Károly Főiskola, 2014.
- [29] Z. USISKIN: *Beauty and serendipity in teaching mathematics*, Educational Designer 3.12 (2019), URL: <https://www.educationaldesigner.org/ed/volume3/issue12/article47/>.