IRSTI 14.35.09
DOI: https://doi.org/10.47344/m8d06v22

Akniyet Orazkhan[1], Nurdaulet Shynarbek[2], Arkhat Baimukhanov[3]*
[1,2]SDU University, Kaskelen, Kazakhstan
[3]Zhetysu University named after I.Zhansugurov,Taldykorgan,Kazakhstan
[*]e-mail: akniyet.orazkhan@sdu.edu.kz

COUNTERINTUITIVE PROBABILITY TASKS IN PRESERVICE MATHEMATICS TEACHER EDUCATION: REASONING, MISCONCEPTIONS, AND RESPONSES TO STUDENT ERROR

Abstract. The current study examines preservice mathematics teachers' answers, explanations, and evaluations of a hypothetical student's errors on six counterintuitive probability tasks. Three types of performance were assessed: selecting an answer, mathematically explaining the chosen answer, and analyzing a hypothetical student's error. Participants were 172 mathematics preservice teachers recruited through convenience sampling from two teacher training institutions in Kazakhstan – a primary sample of 145 from a private university in the Almaty region and a second-site sample of 27 third-year students from another private institution in the Zhetysu region. The research design involved selected-response analysis as well as coding of 1,032 written explanations and 1,032 evaluations with the optional reflections of participants' justifications for their answers and/or critiques of the hypothetical student's solutions. Quantitative analyses addressed task difficulty, block-level comparisons, comparison of course-exposure subgroups in the primary sample, as well as estimation of population-averaged logistic regression models with responses nested within participants. Qualitative analyses included task-based coding of explanations and evaluations, followed by expert audit of the coding scheme. Three main findings are reported. First, the split between Tasks 1–3 and Tasks 4–6 was particularly clear in the primary sample. Second, regardless of location, answer accuracy exceeded the quality of written explanations and pedagogical evaluations because many participants picked the right answer but could not explain why or properly evaluate a hypothetical student's error on the hardest tasks. Third, the second-site sample showed a relatively low task difficulty gradient resulting in a significant site by task block interaction effect.

Keywords: counterintuitive probability tasks, preservice mathematics teachers, probabilistic reasoning, misconceptions, pedagogical content knowledge, student error analysis

Introduction

Despite being structured according to logical principles, mathematics often seems like a complex discipline for people without proper background. However, it does not mean that all the ideas proposed in mathematics are easy to accept. There are several problems which appear simple yet prove to have incorrect first solutions. That is why counterintuitive problems play an important role as they show the contrast between plausible solutions and mathematics-based approaches. A student starts with a certain level of assurance, hesitates after making mistakes, and finally arrives at a more profound understanding of the structure of the problem.

Preservice mathematics teachers should solve the task properly, explain why wrong solutions emerge, and help learners to move from intuitive to mathematical reasoning (Elbehary, 2021; Hokor et al., 2021). At that, merely knowing the right answer is not enough. When children go to schools, they already know something about probability, fair events, numbers, patterns, etc. A teacher who fails to separate intuitive and mathematical thinking will not be able to discuss learners' opinions.

It seems particularly important in probability since the most of its basic principles contradict common sense (Laura et al., 2017; Knudtzon, 2019). Events perceived by people as less random have the same probability as other events. No streaks change the probabilities of outcomes when it comes to an independent process. Some rare conditions can result in lots of false positives. Samples tend to be more variable when they are small.

Teacher education should pay particular attention to this point. Preservice mathematics teachers stand at the boundary between being learners and becoming instructors. Their level of understanding affects how they explain concepts, design examples, respond to questions, and interpret student mistakes. If they can produce a correct answer while failing to understand why intuition misleads, classroom teaching may become procedural rather than explanatory (Burgos et al., 2022a, 2022b; Demirci et al., 2017).

Although considerable attention is paid to probabilistic reasoning and teacher learning, counterintuitive tasks are not well investigated in the field of teacher education (Álvarez-Arroyo et al., 2024; Hokor et al., 2021). Many teacher training programs emphasize formulas and algorithms rather than tasks which require revising one's initial intuitions. Thus, there is reason to suppose that pre-service teachers have insufficient experience in working with misconceptions that appear when probability is at odds with common sense. There is no doubt about both practical and scientific necessity to investigate them more thoroughly.

Against this backdrop, it is necessary to view probability tasks not only in terms of correct and incorrect solutions. In fact, they can also function as diagnostic tools indicating patterns of reasoning, repeated mistakes and the level of pedagogical preparedness when dealing with student errors (Korkmaz & Alkan, 2023). Hence, the current study is focused on three related aspects of students' performance including the correctness of their solution, the correctness of the provided mathematical explanation and the correctness of the evaluation of their peer's mistake. Thus, this study will go beyond assessing preservice teachers' mathematical knowledge. Instead, the research will focus on the relationship between mathematical understanding and readiness for teaching mathematics.

Research Questions

1. How successfully do preservice mathematics teachers solve counterintuitive probability tasks?
2. What types of reasoning do preservice mathematics teachers use when justifying their answers to counterintuitive probability tasks?
3. What misconceptions appear most frequently in preservice mathematics teachers' responses to these tasks?
4. How effectively do preservice mathematics teachers identify and respond to hypothetical student misconceptions related to these tasks?
5. What is the relationship between success in solving counterintuitive probability tasks and the quality of pedagogical responses to student error?

Literature Review

Indeed, probability has remained one of the areas where students struggle the most within mathematics education due to the inherent contradiction between probability and common sense. In many cases, students might have the necessary visual structures or methods to solve certain mathematics problems. However, probability offers additional challenges because things that seem to happen infrequently actually occur frequently, results of a random process do not seem to be random, and decisions that would seem to be fair or logical might lead to the wrong solutions. Probability problems tend to be highly counterintuitive, and thus people often make mistakes while trying to apply common-sense reasoning to these issues (Knudtzon, 2019; Laura et al., 2017). Therefore, probability has long been associated with a domain in which reasoning diverges from intuition significantly. Moreover, such problems shed light on the core of probability thinking and reasoning – at first glance, people's decisions may be guided by experience, resemblance, or balance, rather than underlying mathematics structures.

Such an issue is particularly important when it comes to training future teachers of mathematics. Apart from solving probability tasks correctly, prospective teachers need to explain their solutions, find an explanation for learners' mistakes, and offer the necessary corrective actions. Hence, the training of such specialists should focus not only on helping them solve tasks correctly. Instead, they need to learn how to understand the moment when their intuitions help or mislead them. It turns out that prospective teachers' ability to explain student misconceptions and provide adequate solutions is more important than their knowledge of probability problems. Hence, teacher training programs in probability must concentrate on prospective teachers' interpretation of students' reasoning (Burgos et al., 2022a, 2022b; Bursalı & Özdemir, 2019; Korkmaz & Alkan, 2023).

An extensive corpus of empirical work reveals that preservice teachers are susceptible to a number of probabilistic biases typically exhibited by secondary school students and adults. Several groups of misunderstandings consistently emerge from research. The representativeness bias is one of the most common: in this case, individuals evaluate the likelihood of a certain event based on whether or not it looks like their representation of randomness. As Blanco and Chamberlin (2019) and Hokor et al. (2021) have found, mixed sequences of events are often perceived as more probable than clustered ones in the cases when all the outcomes have identical chances to occur. One of the other typical biases is the gambler's fallacy: individuals believe that random events tend to self-correct over a brief period and predict a reversal after several successful attempts. In this respect, Hokor et al. (2021) point out that even preservice teachers display this behavior pattern because formal training fails to remove intuitive expectations.

In addition, preservice teachers suffer from the problem of equiprobability bias, assuming that any event which they can see is equally probable regardless of whether there is no uniformity in the sample space. According to Dayal and Sharma (2020) and Hokor et al. (2021), preservice teachers use symmetry to determine the odds of an event without counting possible outcomes systematically. There is also some data that suggests that preservice teachers encounter challenges associated with probability when it has to be interpreted through more complicated means. Thus, Batanero et al. (2015) demonstrated that preservice teachers faced semiotic conflicts while solving problems relating to calculating simple, joint, and conditional probabilities using two-way tables. Overall, Kuzu and Arıcan (2020) show that there are inconsistent profiles of competence in statistics and probability among preservice teachers.

The complexity is further compounded where probability problems require knowledge of conditional statements, previous data, or variations within different samples. According to Köklü (2024), many preservice teachers demonstrate poor comprehension of Bayesian and conditional inference, especially with regard to the combination of base rate information with diagnostic tests. The same challenges have been identified by Álvarez-Arroyo et al. (2024), where people struggle to comprehend probabilistic arguments in media contexts, including conditional probability inversion and the misinterpretation of correlations as causality. These similar findings were also reached by Kuş and Çakıroğlu (2020) on the basis of news stories. Overall, these findings show that probability mistakes extend beyond theoretical exercises and exist in socially relevant scenarios despite the prior exposure of learners to concepts and formulas for solving such problems. This highlights that Batanero and Álvarez-Arroyo (2023) call for a new approach in teaching probability.

Moreover, it emerges from the literature review that having knowledge of probability does not necessarily mean being able to teach it. Studies focusing on pedagogical content knowledge reveal that future teachers can recognize the incorrectness of students' responses, but cannot explain their reasoning and suggest an appropriate way to teach them. According to Burgos et al. (2022a, 2022b), preservice teachers usually recognize the superficial nature of mistakes without realizing their underlying reasoning. Korkmaz and Alkan (2023) report similar findings: participants could identify an incorrect answer but struggled to explain the underlying concepts, such as independence, sample space structure, or compound event comparison. The same deficiency can be observed in the process of evaluating students' knowledge of conditional probability, according to Dröse et al. (2022). He (2025) suggests that conventional methods, like tree diagrams, are not sufficient to determine if preservice teachers have grasped the meaning of students' mistakes. These results are important for the discussion since teaching mathematics requires knowledge of content and the ability to use learners' mistakes as evidence of their reasoning. Moreover, this assumption is supported by the theoretical basis. Elbehary (2022) claims that professional knowledge related to probability needs to be rethought as a type of probabilistic reasoning, not only as mastery of procedures and discrete content knowledge. In this context, it is necessary for teachers' knowledge in probability to comprise three aspects, namely reasoning within probability, reasoning about probability language, and interpreting probability notions in educational settings.

Contemporary research has increasingly focused on diagnostic tests, tasks, and response interpretation measures. According to López-Martín et al. (2025), by requiring prospective teachers to interpret student responses, one is able to get information about competencies which are otherwise not detectable using typical test items. Additionally, Drollinger-Vetter et al. (2025) prove that interventions aimed at structured development of pedagogical content knowledge for teaching probability enable professional growth; however, their effect cannot always be considered uniform depending on intervention type used. Alonso-Castaño et al. (2021) note that although preservice teachers may have no difficulties in solving probability problems, their ability to generate the same problems or evaluate another student's approach might be limited. The reviewed papers highlight three distinct yet interconnected aspects: choosing the correct solution, providing mathematical justification, and assessing another person's reasoning. Literature analysis clearly demonstrates that such competencies cannot be equated to each other. In terms of instructional research, Hokor (2020) demonstrates the effectiveness of a constructivist approach in helping teachers overcome misconceptions regarding probability concepts. Intervention-oriented studies show that teacher education could benefit from systematic efforts aimed at developing informal inferential reasoning and pedagogical content knowledge (Vetten et al., 2017). Systematic review research shows that games offer unique opportunities for probability learning at school and teacher-training level (Sharma et al., 2021). Finally, recent work on training cycle proves that competence development could be the foundation of probability-focused teacher education (Su et al., 2024).

Notwithstanding advances achieved in research in both domains, certain limitations of existing studies remain. For example, a large number of studies focused exclusively on the first aspect of competence – answer correctness or content – giving insufficient attention to reasoning and even less attention to hypothetical student reasoning evaluation. Second, while the instruments used to assess the two aspects mentioned above are abundant, a considerable number of assessments measure only one kind of misconception instead of examining counterintuitive probability structures as whole. Finally, although numerous biases have been described, very few researchers considered the connection between mathematical performance and diagnosing learner thinking. It is important to note that a preservice teacher might come up with a correct answer to a task despite using faulty logic, thus failing to provide an explanation of why the given answer makes sense. In other words, without addressing both types of competence simultaneously, one can significantly overestimate teachers' readiness for the class.

The current study seeks to fill this research niche, adopting counterintuitive probability tasks as multi-faceted diagnostic tools in preservice teachers' education. Unlike previous studies, this paper examines three components of competence: answer correctness, written probabilistic reasoning, and hypothetical reasoning evaluation. This makes it possible to identify persistent misconceptions and assess whether correct answers are supported by adequate conceptual understanding. Hence, the present study adds another brick to the literature on probability misconception diagnosis and pedagogical assessment. The novelty of this research lies in the application of a dual diagnosis approach wherein six tasks are used to contrast correct multiple-choice answers, probability explanations, and the identification of a potential student mistake. In doing so, this study is able to provide insights into how correct answers may conceal inadequate conceptual understanding and insufficient pedagogic diagnosis, an observation that would be difficult to make in other studies examining only the correctness of responses.

Methods and Materials

The present research adopted an integrated approach to the mixed methods study of a diagnostic test instrument. Quantitative and qualitative data collection occurred simultaneously using one written instrument. Nevertheless, data analysis was conducted in sequence. Selected responses were first analyzed for task difficulty, block-level differences, course exposure variation, and relationships among performance measures. This led to further qualitative examination of written rationales and evaluations, especially focusing on the most difficult tasks and those where a correct choice was complemented with a poor rationale. Thus, the study features both mixed evidence sources and an integrated design with sequential analysis of data obtained with a single instrument (Elbehary, 2021; Hokor et al., 2021; Korkmaz & Alkan, 2023).

Study participants included preservice mathematics teachers studying at two mathematics teaching programs in Kazakhstan. In the main study group there were 145 students from a private institution located in the region of Almaty. As far as the distribution of the respondents according to the grade helped in analyzing the differences in course exposure, the second site extension study involved 27 third-year students from another private institution in the Zhetysu region. Convenience sampling was used at both study locations because the study required selecting intact groups participating in regular teaching program activities.

Data were gathered through an instrument involving six counterintuitive probability questions. The questions involved topics such as random sequences, gambler's fallacy, outcome counting, conditional probability with base rates, compound events without replacement, and sample-size reasoning. In each question, participants had to give three answers: a choice among the options provided, an explanation for their choice, and an assessment of the reasoning underlying another student's response. Hence, the instrument aimed to collect more information than simply whether the responses were correct, including reasoning processes and diagnostic evaluations of potential misunderstandings, consistent with the literature on the topic (Batanero et al., 2021; Dröse et al., 2022; He, 2025; Knudtzon,2019).

In the quantitative analysis strand, attention was paid to chosen answers and patterns at the task level. Specifically, the analyses covered item-level correctness, within-participant block differences, and pre- versus post-course comparisons within the Almaty-region sample. Population-averaged logistic regressions with responses nested within participants were also estimated. For the multi-site analysis, institutions were compared both descriptively and on a year-matched basis. The Year 3 subgroup from the Almaty-region institution served as the most appropriate reference for the entirely third-year Zhetysu-region sample. The six-task combination served as a diagnostic tool, given that it assessed different misconception families.

The qualitative analysis considered the textual information. The justifications and evaluations were analyzed within the context of the task question, participant selection, and a hypothetical student response. This approach facilitated the study's explanation of the noticeable separation of Tasks 1–3 from Tasks 4–6, the typical difference between correct responses and sound reasoning, and the low instructional diagnostic utility found in the difficult tasks.

The accuracy counts included blank responses as errors. Blank responses that occurred in justifications and evaluations were retained as distinct qualitative categories. Participant identifiers were stripped from the data set and assigned numerical case numbers. Participants submitted their responses during normal risk-free program activities through research documents without identifying information. Participation was voluntary, informed consent was obtained, and anonymity was preserved during data processing and reporting.

Context and analytic procedure

Participants were enrolled from two different institutions. The first one was a private university located in the Almaty region, which included students ranging from Year 1 to Year 5. Second, the participants of the comparison group came from the private university located in the Zhetysu region, which consisted of 27 students of Year 3. In relation to the analytical part of the sample in the Almaty region, it was divided into the pre probability course group and post probability course group. The pre-course group consisted of Years 1 and 2, while the post-course group consisted of Years 3 and 4, and had already taken the probability course. One participant of Year 5 from the Almaty region was used in the descriptive statistics part but was excluded from comparisons. As a result of only having Year 3 students in the Zhetysu region, comparisons between institutions were conducted in terms of the whole sample from the Almaty region and also a subgroup of Year 3 from the Almaty region sample.

With regard to the analysis of selected-response questions, the key focus of analysis was set around the answer to one fundamental question about stability in six tasks. While the first three tasks dealt with exact sequence reasoning, independence, and visible counting, the last three tasks examined base rates, structure of hidden compound events, and variability in sample size. Consequently, the analysis included description, comparison between two task blocks within a student, comparisons between groups exposed to the course at the level of each task, and population-averaged logistic regression with responses nested within individuals.

Both the 1,032 written explanations and the 1,032 evaluative responses were coded separately utilizing a task-based coding scheme. During the first coding phase, an extensive code reflecting the dominant reasoning pattern was attributed to each response. During the second coding phase, the task-based code was appraised using a set of general criteria and boundary rules with specific emphasis on cases when there is a right answer but a weak explanation. The resulting coding scheme allowed for discrimination between good, partial, misconception-driven, empty, and irrelevant written justifications and also between substantive, weak, wrong, confusing, and empty evaluation of the hypothetical learner. After the preliminary coding was complete, the codebook and coded matrices were reviewed by two experts in mathematics education as a coding audit.

Results

The results are organised in three layers. First, the main findings are reported for the primary Almaty-region institution sample, which remains the analytic base because it contains the full year distribution. Second, the Zhetysu-region institution sample is introduced as a second-site extension. Third, the two sites are compared directly, with year-matched analyses used where appropriate.

Table 1

Participant Profile by Institution

Institution	n	%
Private institution in the Almaty region	145	84.3
Private institution in the Zhetysu region	27	15.7
Total	**172**	**100.0**

Within the Almaty-region institution sample, Years 1 and 2 formed the pre-course group (n = 68), and Years 3 and 4 formed the post-course group (n = 76). The Zhetysu-region institution sample consisted entirely of third-year students. This imbalance makes the year-matched Almaty-region institution Year 3 comparison especially important for institutional interpretation.

In the primary Almaty-region institution sample, the mean selected-response score was 3.88 out of 6, with a standard deviation of 1.18 and a median of 4. Only 12 respondents answered all six items correctly. Internal consistency for the six selected-response items was low (Cronbach's $\alpha = 0.23$), which is consistent with the interpretation of the instrument as a diagnostic battery probing distinct misconception families rather than as a single scale.

In the primary Almaty-region institution sample, task-level performance showed a sharp split. The first three tasks were answered correctly by large majorities of students, whereas the last three tasks were much more difficult. Nearly two thirds of the participants, 64.8%, solved all three foundational tasks correctly, but only 13.1% solved all three advanced tasks correctly. A Wilcoxon signed-rank test confirmed that foundational block scores were significantly higher than advanced block scores ($p < .001$).

Table 2

Task-Level Selected-Response Performance in the Primary Almaty-Region Institution Sample

Task	Core idea	Correct (n)	Correct (%)	Item–rest *r*
1	Random sequence	123	84.8	0.07
2	Gambler's fallacy	129	89.0	0.15
3	Sum of two dice	117	80.7	0.14
4	Conditional probability and base rate	64	44.1	0.23
5	Without replacement	66	45.5	0.04
6	Sample size	63	43.4	0.03

This primary-sample pattern was replicated in a population-averaged logistic model with responses nested within students. Correctness was far more likely on the foundational block than on the advanced block (odds ratio = 7.80, z = 7.97, p < .001). Course exposure showed no significant main effect (z = 0.50, p = .621), and there was no significant interaction between course exposure and task block (z = -0.51, p = .611). In other words, the dominant divide in the primary sample was task type rather than year level or probability-course exposure.

Table 3

Course-Exposure Comparison for Selected Answers in the Primary Almaty-Region Institution Sample

Measure	Pre-course	Post-course	Statistic	*p*
Total score, mean	3.84	3.88	$t = -0.22$	0.826
Task 1 correct, %	88.2	81.6	$\chi^2 = 1.23$	0.268
Task 2 correct, %	86.8	90.8	$\chi^2 = 0.59$	0.443
Task 3 correct, %	80.9	80.3	$\chi^2 = 0.01$	0.925
Task 4 correct, %	45.6	42.1	$\chi^2 = 0.18$	0.674
Task 5 correct, %	48.5	42.1	$\chi^2 = 0.60$	0.439
Task 6 correct, %	33.8	51.3	$\chi^2 = 4.48$	0.034

Within the Almaty-region institution sample, the comparison by course exposure produced a restrained picture. Students who had not yet taken the probability course had a mean total score of 3.84, whereas students who had taken the course had a mean total score of 3.88. The difference was trivial and not statistically significant. At the item level, only Task 6 showed an unadjusted post-course advantage. Even there, the broader pattern was persistence rather than broad remediation of counterintuitive misconceptions.

Second-site extension and cross-university comparison

The second sample consists of twenty-seven students who belong to the third year of the preservice teacher training in mathematics. Considering the composition of the second sample which consists of students of the third year, the most relevant comparison for such research will be the third-year students of an educational institution in the Almaty region.

Table 4
Task-Level Selected-Response Accuracy Across Institutions and Year-Matched Comparison

Task	Almaty-region institution total %	Almaty-region institution Year 3 %	Zhetysu-region institution %	*p* (Zhetysu-region institution vs Almaty Year 3)
T1	84.8	82.4	59.3	0.026
T2	88.3	88.2	88.9	0.932
T3	81.4	82.4	70.4	0.223
T4	44.8	43.1	66.7	0.048
T5	45.5	41.2	59.3	0.128
T6	44.1	56.9	59.3	0.838

The second sample did not simply replicate the pattern observed in the first sample. The students who came from the institution in Zhetysu region showed substantially poorer results in Task 1, very similar results in Task 2, slightly worse results in Task 3, but better results in Tasks 4 to 6. As far as the task-level analysis between the two groups is concerned, the greatest discrepancies are in lower scores in Task 1 ($p = .026$) and higher scores in Task 4 ($p = .048$).

Table 5
Block-Level Comparison and Institutional Interaction

Comparison	Tasks 1 to 3 mean	Tasks 4 to 6 mean	*p* (Within-site contrast)
Almaty-region institution total	2.54	1.34	$p < .001$
Almaty-region institution Year 3	2.53	1.41	$p < .001$
Zhetysu-region institution	2.19	1.85	0.179

In block terms, the contrast that dominated the Almaty-region institution sample was much flatter in the second site. The primary Almaty-region institution sample averaged 2.54 out of 3 on Tasks 1 to 3 and 1.34 on Tasks 4 to 6, whereas the Zhetysu-region institution averaged 2.19 and 1.85, respectively. In a year-matched population-averaged logistic model, the decline from the foundational block to the advanced block was significantly smaller in the Zhetysu-region institution than in the Almaty-region institution Year 3 subgroup (site × block odds ratio = 3.64, $p = .008$). This means that the second site preserved the general value of the instrument but not the exact same internal difficulty structure.

Analysis of the profiles of students revealed consistency of results across different locations. For the sample of an institution located in the Almaty region, there was a predominance of the front-loaded type of difficulty distribution, which showed that success tended to occur in Tasks 1-3 but then decreased significantly in Tasks 4-6. The institution situated in the Zhetysu region had more profiles which demonstrated success across the board and also had late strength.

Written justifications and misconception patterns

The verbal reasoning demonstrated a lot of weaknesses compared to those found in the selection responses. Out of the total 870 explanations provided, only 233 were coded as valid while 286 were partial explanations. Another 282 were misconceptions and 69 were not filled. The strongest reasoning was in Task 1 and Task 2 while the weakest was in Task 4, 5, and 6.

Table 6
Quality of Written Justifications by Task in the Primary Almaty-Region Institution Sample

Task	Sound (n)	Sound (%)	Partial (n)	Misconception (n)	Blank/Unclear (n)
1	59	40.7	51	29	6
2	84	57.9	38	14	9
3	51	35.2	54	33	7
4	5	3.4	29	92	19
5	19	13.1	71	43	12
6	15	10.3	43	71	16

A clear difference arose between option selection and justification on the more difficult questions. On question four, sixty-four respondents chose the right response, while only five gave a valid justification for the answer. Question five was answered correctly by sixty-six participants, but only nineteen were able to give a valid justification for their choice. Question six had sixty-three correct answers, but only fifteen could give a valid justification.

Table 7
Dominant Misconception Patterns with Illustrative Translated Excerpts

Task	Dominant misconception	Illustrative translated excerpt	Illusttive sound reasoning
1	Representativeness bias	The more mixed sequence looks more random, so it is more likely.	Each exact sequence has probability 1/64, so the two sequences are equally likely.
2	Gambler’s fallacy	Tails is more likely now because heads has happened too many times.	Each toss is independent, so the sixth toss remains one half and one half.
3	Outcome-counting error	The sums 6 and 7 are equal because the possibilities are the same.	There are six ordered pairs for 7 and five ordered pairs for 6, so 7 is more likely.
4	Base-rate neglect	The test is 90% accurate, so a positive result means the person probably has the disease.	The disease is rare, so false positives outnumber true positives.
5	Single-path reasoning	Red then blue is $3/5 \times 2/4$, so it is equal to two reds.	One red and one blue has two orders, red-blue and blue-red, so the probabilities must be added.
6	Sample-size neglect	The larger hospital is more likely because more children are born there.	Extreme percentages occur more easily in the smaller sample.

Evaluating the hypothetical student’s reasoning

It was found that the evaluative responses pointed out another shortcoming in the pedagogy. Of the 870 evaluative responses, 192 were mathematically grounded, 280 were correct but partial, 188 adopted an incorrect stance, 51 were ambiguous, and 159 were blank. It seems that tasks 4, 5, and 6 revealed poor evaluative skills.

Table 8

Quality of Evaluative Responses to the Hypothetical Student by Task in the Primary Almaty-Region Institution Sample

Task	Grounded correct (n)	Correct but thin/mixed (n)	Incorrect stance (n)	Unclear (n)	Blank (n)
1	48	58	9	6	24
2	50	53	13	7	22
3	33	67	14	10	21
4	23	23	56	9	34
5	12	47	49	9	28
6	26	32	47	10	30

Taken together, the results show that content performance and written competence were related but far from identical. The strongest descriptive summary of that mismatch is the task-level contrast among selected answers, sound justifications, and grounded pedagogical evaluations.

Figure 1 compresses this pattern into a single visual display and makes the sharp decline in explanation quality and grounded pedagogical evaluation on Tasks 4 to 6 immediately visible.

Figure 1

Accuracy, explanation, and pedagogical diagnosis by task in the primary Almaty-region institution sample.

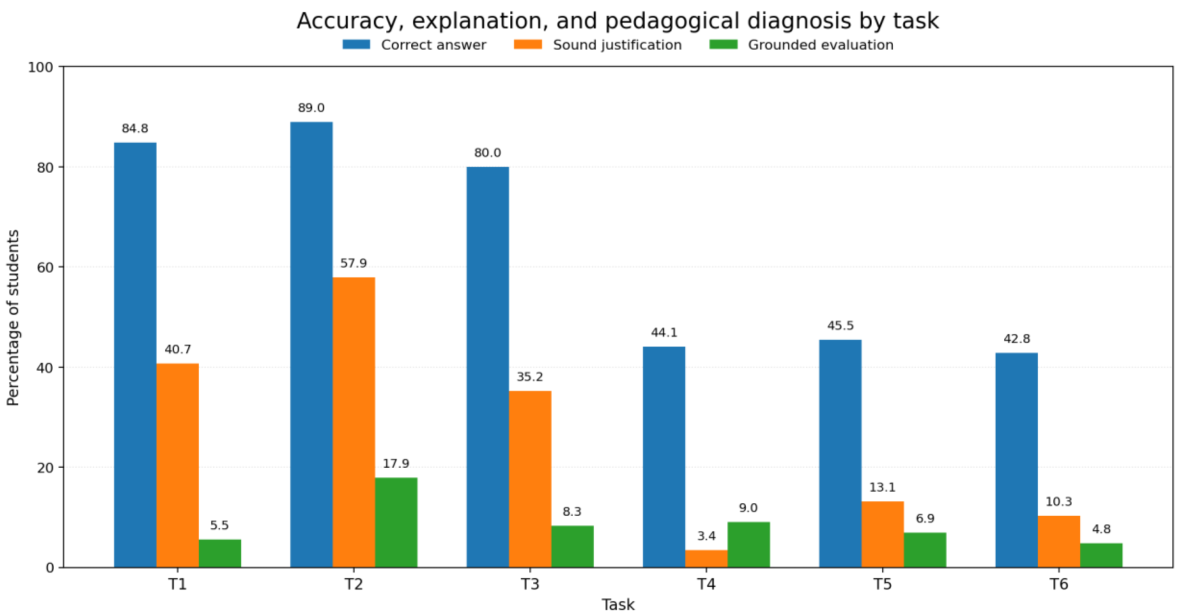


Table 9

The Answer-Reasoning Gap in the Primary Almaty-Region Institution Sample

Task	Correct Answers (n)	Sound Justifications (n)	Sound Among Correct (%)	Grounded Evaluations (n)
1	123	59	48.0	48

2	129	84	65.1	50
3	117	51	43.6	33
4	64	5	7.8	23
5	66	19	28.8	12
6	63	15	23.8	26

Qualitative elaboration of the hardest tasks

The highest deviation between picking the right option and explaining the chosen response is shown in Task 4. The misconception-driven answers highlighted the high precision of the test (90%) and equated this statistic with the probability of disease presence in the case of a positive test result. In turn, better answers focused on comparing the predicted number of true and false positives and included the low prevalence of the disease in their reasoning.

The most evident linear thinking was seen in Task 5. Several students evaluated only one path through which the mixed colour combination could take place, despite there being two possible paths in this case. The best explanations pointed out the two paths (red/blue and blue/red combinations).

In Task 6, the sample-size effect was largely overlooked. The misconception-driven answers argued that both hospitals were equally likely to show an extreme percentage of boys since there was a chance of 1/2 for a boy birth. The better answers addressed the issue of variability and noted that extremes were more probable in small samples.

As the qualitative analysis of the cases demonstrates, the problem cannot be simply attributed to low academic achievement of the respondents. While many students picked the right answer, they still lacked the skills necessary to formulate the underlying probabilistic principle or to account for the gap between the given example and the learner’s understanding of the problem. This distinction is important in teacher preparation as a good class explanation requires more than choosing the right answer.

Optional reflections confirmed the overall tendency. Out of 101 nonblank answers, Task 4 was perceived as the hardest task by 49 respondents while 27 of the participants chose Task 6 as the most problematic one.

Second-site qualitative extension

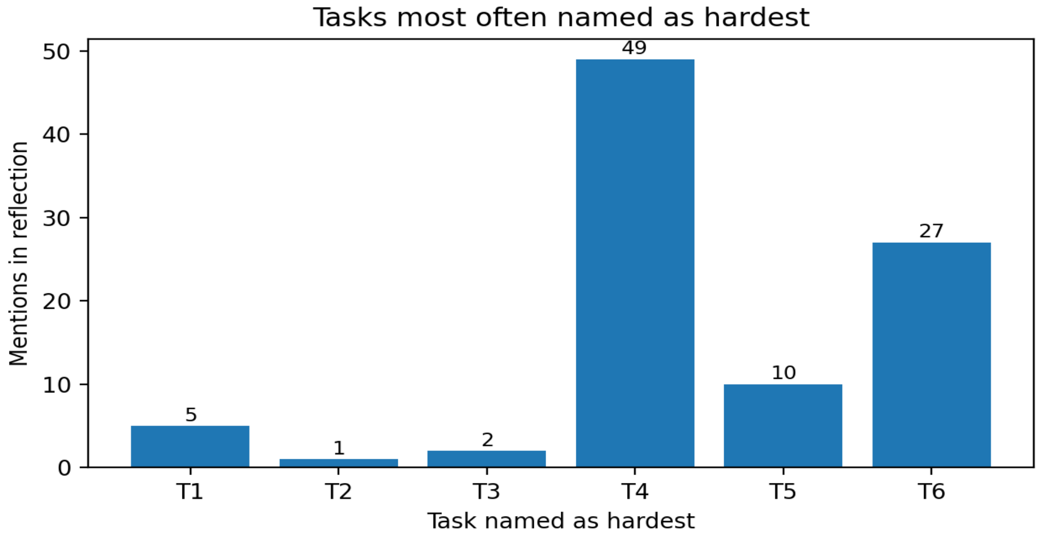
The use of the same coding system to analyze written responses from the institution in the Zhetysu region confirmed the appearance of the very same types of misconceptions in Tasks 1 (representativeness), 2 (gamblers' fallacy), 3 (outcome counting), 4 (base rate neglect), 5 (single path reasoning), and 6 (sample size neglect) as in the first sample. Thus, the second sample provided additional evidence for the conceptual validity of the instrument in different institutions.

Alongside with the similarities in the types of errors committed by students, the qualitative analysis revealed a different distribution of strengths among them. In contrast to the primary sample, students from the Zhetysu region provided relatively frequent justifications of their answers that involved the use of explicit reasoning. For example, several responses to Task 4 compared true and false positive cases; several responses to Task 5 made the two paths in the mixed colours event apparent; and some responses to Task 6 referred to greater variance in smaller samples. This pattern is consistent with stronger probabilistic reasoning on these particular tasks, although the small second-site sample prevents strong institutional claims.

It is worth pointing out, however, that the difference between the samples was not absolute but was confined to some tasks only. As seen from Task 1, assertion-level agreements with the equal probability claim prevailed in the secondary sample. As can be seen, the qualitative results confirm the quantitative findings and reinforce the interpretation that the institutional differences between two samples are profile differences.

Figure 2

Tasks named as hardest in the optional reflection in the primary Almaty-region institution sample.



Discussion

In light of the presented results, one may suggest that the engagement of preservice mathematics teachers with the counterintuitive probability problems should not be understood only in terms of accurate vs. inaccurate answers. The collected data show that, for quite a number of participants, it was possible to choose a correct option, but not to justify it adequately. Moreover, when asked to analyze hypothetical student reasoning, a majority of participants was able to point out that something was wrong about the solution, but not able to specify this issue in terms of mathematical concepts and principles. It would be fair to assume that there is an obvious relationship between the ability to provide a correct answer, the skill of probabilistic reasoning, and pedagogical diagnosis, but these skills should not be confused as one kind of competence.

One of the most evident conclusions that can be drawn from the results obtained is the inequality in the level of difficulty of all six tasks tested. While three first tasks were completed successfully by the vast majority of participants (random sequence, gambler's fallacy, sums of two dice), three other problems appeared to be much harder (base rate, without replacement, sample size). This conclusion is important since it indicates that the counterintuitive probability tasks do not belong to the same level of cognitive activity. Some questions may represent familiar types of intuitive errors, whereas other tasks presuppose more profound control of conditional structure, event composition, and variability of samples. In other words, the presented results prove the findings made in previous researches concerning the complexity of probability understanding.

In addition, the results indicate that correct answers do not necessarily reflect participants' knowledge of probabilistic concepts. This gap was most evident on the complex tasks — base rate, without replacement, and sample size. Participants chose the correct answer but could not explain the role of prevalence, sensitivity, or false positives. It can be concluded that the presented problem is educationally important since it refutes the generally accepted opinion regarding teacher preparation programs which imply that the correct answer is synonymous with adequate knowledge.

Misconception patterns in the written responses correspond well to previous findings. For example, representativeness reasoning was frequently the cause of the wrong choice in the random-sequence problem, where mixed sequences were viewed as random and therefore seen as having higher probability. Participants who took the gambler's-fallacy problem showed a subset of beliefs that tails became more probable after several successes (heads). For the dice-sum problem, errors in counting all possible outcomes as well as equating probabilities of all observable outcomes regardless of the structure of sample space were prevalent sources of misconceptions. The base-rate problem displayed a strong misunderstanding, as many participants focused on the accuracy of the testing while ignoring its low prevalence. Without-replacement problem displayed inability to consider all relevant paths, and sample-size problem revealed that many participants were unaware of the fact that extreme proportions happened more often in small sample sizes.

Furthermore, the present findings contribute to the literature by showing that misconceptions can persist despite effective solving. Specifically, a significant challenge was in evaluation of hypothetical student reasoning. First of all, some participants gave the right answer to the question and argued it briefly. Second, others stated the right solution but failed to provide an analysis of the student's reasoning. Finally, there were also those who agreed with the student based on an initial look at the answer. Thus, the findings imply that pedagogical content knowledge of probability is more complex than mere content knowledge. Effective probability teaching needs future teachers to understand what makes intuitive explanations compelling and how to guide students toward mathematical reasoning.

Moreover, comparing students who had not yet taken a probability course with those who had already taken one is also essential. Indeed, in this case, we cannot speak about causation and thus should not make any strong claims. However, still, this information can be used for our interpretation since the association with formal course exposure appeared limited in terms of solving counterintuitive problems. Specifically, only one of the examined problems showed the post-course advantage, which cautiously suggests that routine instruction alone may not remove deeply rooted intuition-related misconceptions.

The secondary-site comparison provides an additional contribution to the literature. The latter does not simply illustrate the same pattern in the context of a broader aggregated sample. Instead, the secondary comparison helps examine the diagnostic usefulness of the instrument and simultaneously suggests differences in the prominence of the various misconception families among institutions. While in the sample from the Almaty region, the most common pattern was characterized by a sharp drop from the first three to the last three items, in Zhetysu, the decrease was much less pronounced, implying relatively high results in solving base-rate, without-replacement, and sample size problems and relatively low scores for the random sequence item. This descriptive pattern is important because it suggests the presence of stable counterintuitive challenges, along with those that depend on particular educational circumstances.

Another point of reflection concerns the distinction between mathematical thinking and professional readiness. There is the likelihood that preservice teachers could correctly solve the problems, while being unable to deal with schoolchildren's erroneous answers caused by misconceptions. The conclusion is that the development of diagnostic competence is essential in teacher training programs and should be done independently of mathematical problem-solving skills.

Therefore, the discussed study has several implications both from theoretical and practical perspectives. From a theoretical point of view, it supports the three-layered model of professional competence in terms of probability, including answer selection, justification, and diagnosis. At the same time, the layers overlap; however, their formation is not the same. Practical implications are related to the role of counterintuitive probability tasks in teacher education programs. Such activities have potential for diagnosing student errors, prompting written justifications, and generating discussion.

Limitations

A few limitations need to be mentioned. Firstly, while the study draws on data from two different institutions, the second site is much smaller and represents only the third year of students, so the institutional comparison is not fully equal. Moreover, the comparison of pre-and post-course samples is cross-sectional and cannot be considered as proof of the causative effect of the intervention. Third, the heterogeneity of the selected response questions used as part of the questionnaire is likely the reason behind the low internal consistency coefficient, or alpha. Finally, the qualitative coding relied on an expert audit rather than independent re-coding of the full response corpus, so formal inter-rater agreement statistics were not computed. Future studies should employ independent parallel coding to strengthen the credibility of qualitative classifications.

Conclusion

This study shows that by using a short set of counterintuitive probability problems, one can assess three different, but interrelated aspects of teachers' knowledge: selecting the right mathematical solution, explaining the probability principle behind the problem, and diagnosing learners' misconceptions. The key finding is not that some tasks were harder than others. Rather, correct answer rates consistently exceeded the depth of conceptual understanding — particularly for tasks involving base rates, compound events, and sample size.

Adding the second site allowed revealing the same general pattern at both universities despite differences in the difficulty hierarchy. One important implication concerns the teaching of probability to future teachers. First, one cannot restrict oneself by measuring correctness only when assessing probabilistic knowledge. One also needs to ask preservice teachers not only which solution is correct, but also why a particular solution appears intuitively appealing and why it is inconsistent with probabilistic reasoning.

Use of AI Tools

AI-assisted tools were used only for language improvement, correction of errors and formatting. All scientific decisions, data analysis, interpretation, and conclusions were made and verified by the authors.

References

Alonso-Castaño, M., Alonso, P., Mellone, M., & Rodríguez-Muñiz, L. J. (2021). What mathematical knowledge do prospective teachers reveal when creating and solving a probability problem? Mathematics, 9(24), 3300. https://doi.org/10.3390/math9243300

Álvarez-Arroyo, R., Batanero, C., & Gea, M. M. (2024). Probabilistic literacy and reasoning of prospective secondary school teachers when interpreting media news. ZDM, 56(6), 1045. https://doi.org/10.1007/s11858-024-01586-8

Batanero, C., & Álvarez-Arroyo, R. (2023). Teaching and learning of probability. ZDM, 56(1), 5. https://doi.org/10.1007/s11858-023-01511-5

Batanero, C., Begué, N., Álvarez-Arroyo, R., & Valenzuela-Ruiz, S. M. (2021). Prospective mathematics teachers' understanding of classical and frequentist probability. Mathematics, 9(19), 2526. https://doi.org/10.3390/math9192526

Batanero, C., Contreras, J. M., Díaz, C., & Sánchez, E. (2015). Prospective teachers' semiotic conflicts in computing probabilities from a two-way table. International Electronic Journal of Mathematics Education, 10(1), 3. https://doi.org/10.29333/iejme/288

Blanco, T. G., & Chamberlin, S. A. (2019). Pre-service teacher statistical misconceptions during teacher preparation program. The Mathematics Enthusiast, 16, 461. https://doi.org/10.54870/1551-3440.1469

Burgos, M., Albanese, V., & López-Martín, M. del M. (2022a). Prospective primary school teachers' recognition of proportional reasoning in pupils' solution to probability comparison tasks. HAL. https://hal.science/hal-03751836

Burgos, M., López-Martín, M. del M., Aguayo-Arriagada, C., & Albanese, V. (2022b). Análisis cognitivo de tareas de comparación de probabilidades por futuro profesorado de Educación Primaria. Uniciencia, 36(1), 1. https://doi.org/10.15359/ru.36-1.38

Bursalı, G. G., & Özdemir, B. G. (2019). Matematik öğretmenlerinin ve öğretmen adaylarının kavram yanılgılarına yönelik öğretimsel açıklamaları: Olasılık konusu. Journal of Computer and Education Research, 7(14), 642. https://doi.org/10.18009/jcer.639384

Dayal, H., & Sharma, S. (2020). Investigating probability concepts of secondary pre-service teachers in a game context. Australian Journal of Teacher Education, 45(5), 91. https://doi.org/10.14221/ajte.2020v45n5.6

Demirci, Ö., Özkaya, M., & Konyalioğlu, A. C. (2017). Öğretmen adaylarının olasılık konusuna ilişkin hata yaklaşımları. Erzincan Üniversitesi Eğitim Fakültesi Dergisi. https://doi.org/10.17556/erziefd.310667

Drollinger-Vetter, B., Buff, A., Lipowsky, F., Philipp, K., & Vogel, S. (2025). Fostering pedagogical content knowledge on probability in preservice primary teachers within formal teacher education: A longitudinal experimental field study. Teaching and Teacher Education, 159, 104953. https://doi.org/10.1016/j.tate.2025.104953

Dröse, J., Griese, B., & Wessel, L. (2022). Prospective teachers' diagnostic judgments on students' understanding of conditional probabilities. HAL. https://hal.archives-ouvertes.fr/hal-03744585

Elbehary, S. G. A. (2021). Reasoning under uncertainty within the context of probability education: A case study of preservice mathematics teachers. Pythagoras, 42(1). https://doi.org/10.4102/pythagoras.v42i1.630

Elbehary, S. G. A. (2022). Rethinking mathematics teachers' professional knowledge for teaching probability from the perspective of probabilistic reasoning: A proposed framework. International Electronic Journal of Mathematics Education, 17(3). https://doi.org/10.29333/iejme/12145

He, S. (2025). How do preservice mathematics teachers analyze and respond to student errors in solving probability problems using tree diagrams? Teaching and Teacher Education, 168, 105250. https://doi.org/10.1016/j.tate.2025.105250

Hokor, E. K. (2020). Pre-service teachers' probabilistic reasoning in constructivist classroom. Pedagogical Research, 5(2). https://doi.org/10.29333/pr/7838

Hokor, E. K., Apawu, J., Owusu-Ansah, N. A., & Agormor, S. (2021). Preservice teachers' misconceptions in solving probabilistic problems. Pedagogical Research, 7(1). https://doi.org/10.29333/pr/11441

Knudtzon, S. H. (2019). Pitfalls and surprises in probability: The battle against counterintuition. HAL. https://hal.science/hal-02411606

Köklü, O. (2024). A multidimensional diagnostic assessment for preservice mathematics teachers' conditional probability and Bayesian reasoning. DergiPark. https://doi.org/10.21031/epod.1574880

Korkmaz, E., & Alkan, S. (2023). The preservice teachers' approaches toward incorrect probability problems. SAGE Open, 13(4). https://doi.org/10.1177/21582440231206605

Kuş, M., & Çakıroğlu, E. (2020). Matematik öğretmeni adaylarının bilimsel araştırma hakkındaki eleştirel düşünme süreçleri: Gazete haberi örneği. Turkish Journal of Education, 9(1), 22-37. https://doi.org/10.19128/turje.605456

Kuzu, O., & Arıcan, M. (2020). Investigating preservice middle school mathematics teachers' competencies in statistics and probability in terms of various variables. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 11(1), 13. https://doi.org/10.21031/epod.562586

Laura, A., Mellone, M., Romano, P., & Tortora, R. (2017). Errors or didactic resources: A teacher education task in the context of probability. HAL. https://hal.science/hal-01927877

López-Martín, M. del M., Burgos, M., & Albanese, V. (2025). Interpretación de respuestas de alumnos en problemas de probabilidad por maestros en formación. Revista Interuniversitaria de Formación del Profesorado, 39(2), 179. https://doi.org/10.47553/rifop.v39i2.108921

Sharma, S., Sharma, S., Doyle, P., Marcelo, L., & Kumar, D. (2021). Teaching and learning probability using games: A systematic review of research from 2010-2020. Waikato Journal of Education, 26(2). https://doi.org/10.15663/wje.v26i2.881

Su, C. S., Díaz-Levicoy, D., & Hsu, C. C. (2024). The impact of a training cycle on statistics and probability for future primary schoolteachers with a gender-focused approach within the framework of education for sustainable development. Eurasia Journal of Mathematics, Science and Technology Education, 20(2). https://doi.org/10.29333/ejmste/14177

Vetten, A. de, Schoonenboom, J., Keijzer, R., & Oers, B. van. (2017). Informal statistical inference and pre-service primary school teachers: The development of their content knowledge and pedagogical content knowledge during a teacher college intervention. HAL. https://hal.archives-ouvertes.fr/hal-01927861

Акниет Оразхан[1*], *Нурдаулет Шынарбек*[2], *Архат Баймуханов*[3]
[1,2]SDU University, Каскелен, Казахстан
[3]Жетысуский университет имени Ильяса Жансугурова, Талдыкорган, Казахстан
[*]e-mail: akniyet.orazkhan@sdu.edu.kz

КОНТРИНТУИТИВНЫЕ ЗАДАЧИ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ В ПОДГОТОВКЕ БУДУЩИХ УЧИТЕЛЕЙ МАТЕМАТИКИ: РАССУЖДЕНИЯ, ЗАБЛУЖДЕНИЯ И АНАЛИЗ ОШИБОК УЧАЩИХСЯ

Аннотация. Данное исследование изучает ответы, объяснения и оценки студентов педагогических специальностей по математике при работе с шестью контринтуитивными задачами по теории вероятностей. Оценивались три типа деятельности: выбор ответа, математическое объяснение выбранного ответа и анализ ошибки гипотетического студента. В исследовании приняли участие 172 студента педагогических специальностей по математике, отобранных методом удобной выборки из двух педагогических учебных заведений Казахстана — основная выборка из 145 человек из частного университета Алматинской области и дополнительная выборка из 27 студентов третьего курса другого частного учебного заведения Жетысуской области. Дизайн исследования включал анализ ответов с выбором варианта, а также кодирование 1 032 письменных объяснений и 1 032 оценок с возможными обоснованиями участниками своих ответов и/или критикой решений гипотетического студента. Количественный анализ охватывал сложность заданий, сравнение на уровне блоков, сравнение подгрупп основной выборки по признаку прохождения курсов, а также построение усреднённых по популяции логистических регрессионных моделей с ответами, вложенными в участников. Качественный анализ включал тематическое кодирование объяснений и оценок с последующей экспертной проверкой схемы кодирования. Представлены три основных результата. Во-первых, разрыв между заданиями 1–3 и заданиями 4–6 был особенно выражен в основной выборке. Во-вторых, независимо от места проведения исследования, точность ответов превышала качество письменных объяснений и педагогических оценок, поскольку многие участники выбирали правильный ответ, но не могли объяснить почему или должным образом оценить ошибку гипотетического студента на наиболее сложных заданиях. В-третьих, в дополнительной выборке наблюдался относительно низкий градиент сложности заданий, что привело к значимому эффекту взаимодействия между площадкой и блоком заданий.

Ключевые слова: контринтуитивные вероятностные задачи, будущие учителя математики, вероятностное мышление, заблуждения, педагогическое содержание знаний, анализ ошибок учащихся

Ақниет Оразхан[1*], *Нұрдәулет Шынарбек*[2], *Архат Баймұханов*[3]
[1,2]SDU University, Қаскелең, Қазақстан
[3]Ілияс Жансүгіров атындағы Жетісу университеті, Талдықорған, Қазақстан
[*]e-mail: akniyet.orazkhan@sdu.edu.kz

МАТЕМАТИКА МҰҒАЛІМДЕРІН ДАЯРЛАУДАҒЫ КОНТРИНТУИТИВТІ ЫҚТИМАЛДЫҚ ЕСЕПТЕРІ: ПАЙЫМДАУ, ҚАТЕ ТҮСІНІКТЕР ЖӘНЕ ОҚУШЫ ҚАТЕЛЕРІН ТАЛДАУ

Аңдатпа. Бұл зерттеу математика пәнінің болашақ мұғалімдерінің алты контринтуитивті ықтималдық есептері бойынша берген жауаптарын, түсіндірмелерін және гипотетикалық студенттің қателерін бағалауын қарастырады. Үш түрлі әрекет бағаланды: жауап таңдау, таңдалған жауапты математикалық тұрғыдан түсіндіру және гипотетикалық студенттің қатесін талдау. Зерттеуге Қазақстандағы екі педагогикалық жоғары оқу орнынан ыңғайлы іріктеу әдісімен жинақталған 172 математика мамандығының студент-мұғалімдері қатысты — Алматы облысындағы жеке университеттен 145 адамнан тұратын негізгі іріктеме және Жетісу облысындағы басқа жеке оқу орнының 27 үшінші курс студентінен тұратын қосымша іріктеме. Зерттеу дизайны жауап нұсқаларын талдауды, сондай-ақ қатысушылардың жауаптарын негіздеуі және/немесе гипотетикалық студент шешімдерін сынауы бойынша 1 032 жазбаша түсіндірме мен 1 032 бағалауды кодтауды қамтыды. Сандық талдау тапсырмалардың қиындығын, блок деңгейіндегі салыстыруларды, негізгі іріктемедегі курстан өткен кіші топтарды салыстыруды, сондай-ақ қатысушылар ішіне кірістірілген жауаптары бар популяция бойынша орташаланған логистикалық регрессия үлгілерін бағалауды қамтыды. Сапалық талдауға түсіндірмелер мен бағалауларды тақырыптық кодтау, одан кейін кодтау схемасын сараптамалық тексеру кірді. Үш негізгі нәтиже көрсетілді. Біріншіден, 1–3 тапсырмалар мен 4–6 тапсырмалар арасындағы алшақтық негізгі іріктемеде айқын байқалды. Екіншіден, зерттеу орнына қарамастан, жауаптардың дұрыстығы жазбаша түсіндірмелер мен педагогикалық бағалаулардың сапасынан жоғары болды, өйткені көптеген қатысушылар дұрыс жауапты таңдағанымен, ең қиын тапсырмаларда неге дұрыс екенін түсіндіре алмады немесе гипотетикалық студенттің қатесін дұрыс бағалай алмады. Үшіншіден, қосымша іріктемеде тапсырма қиындығының салыстырмалы түрде төмен градиенті байқалды, бұл орын мен тапсырма блогы арасында маңызды өзара әрекеттесу әсеріне әкелді.

Түйін сөздер: контринтуитивті ықтималдық есептері, болашақ математика мұғалімдері, ықтималдық ойлау, қате түсініктер, педагогикалық мазмұндық білім, оқушы қателерін талдау

Received 10 April 2026