



6. Мосула Л. М., Криськів Л. С., Кучер Т. В. Досвід впровадження нової навчальної дисципліни «Фізико-хімічний аналіз у створенні ліків» у процесі підготовки майбутніх провізорів. *Медична освіта*. 2019. № 2. С. 31–35. URL: <https://doi.org/10.11603/me.2414-5998.2019.2.10340>

7. Лисенко Т. А., Привалко Е. Г., Зайцева Г. М. Особливості використання міждисциплінарних зв'язків у процесі професійно орієнтованого навчання майбутніх фармацевтів на прикладі предмету «Фізико-хімічний аналіз у створенні ліків». *Вісник освіти та науки*. ISSN 2786-6165 (ONLINE) № 9(15) 2023: <http://ir.librarynmu.com/bitstream/123456789/12335/1>



UDC 378.147:519.2:159.955

[https://doi.org/10.52058/2786-6165-2025-11\(41\)-1524-1541](https://doi.org/10.52058/2786-6165-2025-11(41)-1524-1541)

Sachaniuk-Kavets'ka Natalia Vasylivna Candidate of technical sciences, Associate professor, Vinnytsia National Technical University, Vinnytsia, <https://orcid.org/0000-0001-6405-1331>.

Nykyporets Svitlana Stepanivna the senior English language lecturer, Vinnytsia National Technical University, Vinnytsia, <https://orcid.org/0000-0002-3546-1734>.

DEVELOPING CRITICAL THINKING IN STUDENTS OF TECHNICAL SPECIALTIES THROUGH THE MATHEMATICS OF UNCERTAINTY AND EDUCATIONAL DEBATES IN ENGLISH: AN INTEGRATED EXPERIMENTAL-METHODOLOGICAL MODEL

Abstract. The article proposes an integrated experimental-methodological model that combines probability theory, mathematical statistics and structured educational debates in English to develop critical thinking in students of technical specialties under wartime conditions in Ukraine. The aim is to strengthen the evidential basis of engineering judgements by linking formal probabilistic reasoning with transparent L2 argumentation. The study involves 117 first- and second-year students, allocated by cluster randomisation to an experimental condition and an active control. During an eight-week intervention, the experimental groups study core topics in probabilistic reasoning and risk and then translate these tools into debate formats with explicit theses, assumptions, numerical evidence, counterarguments and refutation.

Effectiveness is evaluated using the SKM-TECH critical thinking composite, the DEBATE-ENG rubric and a prediction battery with Brier scores and calibration curves, complemented by source-checking tasks and brief scales of cognitive load and state stress. ANCOVA and multilevel models show significant advantages of the experimental condition on the overall SKM-TECH index and all subscales, with the largest gains in probabilistic reasoning, calibration and source verification. Brier scores decrease and calibration curves move closer to the bisector without an increase in subjective load, indicating automatization of key operations and successful management of extraneous load. English-language argumentation also improves, though with effects moderated by L2 level. Overall, the findings indicate that integrating probability theory with debates can reduce typical cognitive biases, stabilise reasoning during air-raid



related interruptions and offer a scalable template for critical thinking instruction in technical higher education.

Keywords: probability theory, critical thinking, educational debates, technical higher education, wartime education.

Сачанюк-Кавецька Наталія Василівна кандидат технічних наук, доцент кафедри вищої математики, Вінницький національний технічний університет, м. Вінниця, <https://orcid.org/0000-0001-6405-1331>.

Никипорець Світлана Степанівна старший викладач кафедри іноземних мов, Вінницький національний технічний університет, м. Вінниця, <https://orcid.org/0000-0002-3546-1734>.

ФОРМУВАННЯ КРИТИЧНОГО МИСЛЕННЯ СТУДЕНТІВ ТЕХНІЧНИХ СПЕЦІАЛЬНОСТЕЙ ЗАСОБАМИ МАТЕМАТИКИ НЕВИЗНАЧЕНОСТІ І НАВЧАЛЬНИХ ДЕБАТІВ: ІНТЕГРОВАНА ЕКСПЕРИМЕНТАЛЬНО-МЕТОДИЧНА МОДЕЛЬ

Анотація. Стаття пропонує інтегровану експериментально-методичну модель, що поєднує теорію ймовірності, математичну статистику та структуровані навчальні дебати англійською мовою для розвитку критичного мислення студентів технічних спеціальностей в умовах воєнного часу в Україні. Метою є посилення доказової основи інженерних суджень шляхом поєднання формального ймовірнісного міркування з прозорою аргументацією іноземною мовою (L2). Дослідження охоплює 117 студентів 1-2 курсів, розподілених методом кластерної рандомізації на експериментальну умову та активний контроль. Протягом восьми тижневої інтервенції експериментальні групи опановують ключові теми ймовірнісного міркування та ризику, а далі переносять ці інструменти у дебатні формати з чітко сформульованими тезами, припущеннями, числовими доказами, контраргументами та рефутацією.

Ефективність оцінюється за допомогою композиту критичного мислення SKM-TECH, рубрики DEBATE-ENG та батареї прогнозних завдань із розрахунком Brier score і побудовою калібрувальних кривих, доповнених завданнями з перевірки джерел і короткими шкалами когнітивного навантаження та станового стресу. Моделі ANCOVA та багаторівневі моделі демонструють суттєві переваги експериментальної умови за загальним індексом SKM-TECH і всіма його підшкалами, з найбільшими приростами в ймовірнісному міркуванні, калібруванні та верифікації джерел. Показники Brier score знижуються, а калібрувальні криві наближаються до бісектриси



без зростання суб'єктивного навантаження, що свідчить про автоматизацію ключових операцій і успішне керування екстрагенним навантаженням. Англomовна аргументація також покращується, хоча ефекти модеруються рівнем L2. Загалом результати вказують на те, що інтеграція теорії ймовірності з дебатами здатна зменшувати типові когнітивні хиби, стабілізувати міркування під час переривань, пов'язаних з повітряними тривогами, та пропонувати масштабований шаблон навчання критичного мислення у технічній вищій освіті.

Ключові слова: теорія ймовірності, критичне мислення, освітні дебати, вища технічна освіта, освіта у воєнний час.

The full-scale war of Russia against Ukraine has radically changed the operating conditions of higher education institutions, exacerbating fundamental risks to the continuity of the educational process and to academic quality. Interruptions to learning caused by air-raid alerts, forced relocations, periodic blackouts and an elevated level of situational stress create additional cognitive load for students of technical specialities. Against this background, information threats are intensifying, including technical disinformation, manipulative data visualisation and the automated production of English-language texts that imitate an academic style. For engineering education this represents a direct threat of erroneous decision-making, since the reliability of conclusions depends on the combination of formal rigour of thinking with correct argumentation and communication of results [1,2].

Despite the long-standing tradition of teaching elements of probability theory and mathematical statistics within higher mathematics courses, as well as the development of English-language components in ESP (English for Specific Purposes) disciplines, methodological separation between these learning tracks persists. Students often demonstrate [3] the ability to perform probability calculations, but do not command the English-language means required for the precise formulation of the limits of conclusions, the calibration of confidence and the correct interpretation of evidence. For their part, practices of academic writing and oral argumentation in English do not always rest upon [4] formal criteria of evidentiality and the verifiability of numerical claims. This gap leads to persistent errors of thinking (base rate neglect, categorical statements without intervals, conflation of statistical significance with practical importance) and to a false sense of confidence under conditions of informational pressure.

The problem examined in the article consists in the theoretical and methodological substantiation of an integrated approach that connects probability theory with structured educational debates in English as an instrument for the development of critical thinking. Such an approach should ensure both the formal



verifiability of reasoning and its communicative adequacy in L2, taking into account the stress factors of wartime and the need for rapid but well-founded engineering decisions. The scientific and practical significance of the study lies in creating conditions for a reliable transition from numerical estimates of probability to transparent arguments and decisions that meet the requirements of safety, reproducibility and ethical responsibility in Ukrainian technical education during the war.

Aim and objectives

Aim of the study. To develop and empirically test an integrated model that combines probability theory, elements of mathematical statistics and educational debates in English in order to enhance the evidentiality of reasoning, the calibration of confidence and the quality of English-language argumentation among students of technical specialities.

Research objectives.

1. To construct the theoretical framework of the model that aligns formal procedures of probabilistic inference with discursive practices of argumentation in L2.

2. To design an 8-week instructional intervention module with a clear schedule, roles, debate regulations and sets of engineering cases with a controlled level of uncertainty.

3. To operationalise the target constructs and develop assessment instruments: the SKM-TECH critical thinking composite, the DEBATE-ENG debate rubric, a battery of calibration tasks with Brier score, CHECK-FAST micro-protocols for source verification, scales of cognitive load (Paas) and situational stress, and a short test of L2 competence.

4. To carry out a quasi-experiment with pre-test, post-test and delayed post-test, comparing the experimental condition with an active control according to a clustered group design.

5. To conduct a statistical analysis of the effectiveness of the model using ANCOVA and mixed models, to estimate effect sizes and measurement reliability, and to ensure blind review of written work and transcripts.

6. To examine the trade-off between accuracy and load: to check whether improvements in calibration and evidentiality are achieved without a substantial increase in subjective load and stress.

7. To assess moderators of the effect: L2 level, prior statistical training, instructor-related differences, and the robustness of results across different scenarios of uncertainty.

8. To define criteria of success and practical recommendations for scaling: threshold levels of improvement in the SKM-TECH composite and reductions in Brier score, implementation protocols for higher mathematics and ESP courses,



including procedures for ensuring continuity of learning under conditions of interruption.

Theoretical foundations

Uncertainty in the evidential strength of reasoning is understood as changes in the degree of conviction. The updating of beliefs takes place in accordance with Bayes' rule: the posterior probability of a hypothesis is proportional to the product of the prior probability and the likelihood of the data. In practical terms this means working with frequency representations, likelihood ratios and odds, as well as the explicit formulation of assumptions and the limits of applicability of the model. The calibration of confidence is treated [5] as an instructional goal: to bring subjective probability estimates closer to empirical frequencies and to reduce systematic overconfidence. In this context, overconfidence denotes a stable pattern in which stated degrees of certainty exceed the long-run frequencies with which the corresponding claims turn out to be correct.

The set of parameters that are compatible with the data under a fixed procedure of construction is specified by means of a confidence interval, which serves as an instrument for the correct formulation of the limits of conclusions. Null hypothesis tests are used to control type I and type II errors and to make decisions under constraints of resources and time. In an educational context it is important to distinguish statistical significance from practical importance, integrating into debates and written assignments the requirement to provide effect size estimates and confidence intervals in parallel. This disciplines the transition from numbers to appropriate managerial or engineering decisions.

Any managerial decision is associated with risk. In this context risk is understood as a function of the probability of an event and its consequences. For technical cases expected loss values, risk matrices and sensitivity analysis are used. The explication of uncertainty becomes mandatory: the presentation of alternative scenarios, indicators of measurement reliability, the checking of measurement units and the consistency of data. Such boundaries make it possible to transfer statistical conclusions into the decision space without logical leaps and with traceable responsibility for assumptions. Structured educational debates perform the function of operationalising logical norms in communication. A format with a clear thesis, domain of assumptions, numerical evidence, counterarguments and refutation forces students to transform formal results into transparent, verifiable arguments. Conditions of competition and formal rules create a natural check on consistency: any error in base rates, an incorrect interval or a substitution of measurement units immediately becomes the subject of refutation. The combination of debates with Bayesian transformations and the analysis of confidence intervals forms a habit of explicitly indicating the limits of conclusions and the conditions under which a position should be revised.



It should be noted that working in L2 increases intrinsic and extraneous cognitive load due to reduced automaticity of lexico-grammatical operations and higher demands on working memory. In order not to trade accuracy for overload, the instructional design uses load management: stepwise introduction of terms and symbols, partial (applied) formulations instead of abstract ones, templates of English-language expressions of uncertainty, collocation cards for statistical argumentation, time boxes and micro-breaks. Additionally, worked examples are used to support the transition from a model solution to the independent construction of mathematical models, as well as short instruments for monitoring load and stress that allow the instructor to regulate the pace and depth of processing of the material. The integration of Bayesian updating, interval estimates and testing procedures with debate practice in L2 creates a single learning loop: calculation – interpretation – communication – refutation – revision of beliefs. Each element reduces characteristic thinking errors and at the same time increases the transfer of knowledge to real engineering tasks. The management of cognitive load serves as a condition for the stable operation of this loop under the stress and interruptions characteristic of wartime.

Research design

The study employs a quasi-experimental design with pre-test and post-test and cluster randomisation at the level of academic groups. Groups are divided into two subgroups: an experimental subgroup and an active control subgroup. Randomisation is carried out at cluster level in order to minimise the spillover of instructional practices between students from different groups and to avoid contamination.

The experimental condition involves integrated instruction that combines probability theory, elements of mathematical statistics and structured educational debates in English. Participants master probability calculations, Bayesian updating, confidence intervals, hypothesis testing and the basics of risk assessment, and then transfer these methods into a debate format with a clearly formulated thesis, domain of assumptions, mathematical proofs, counterarguments and refutation. The active control condition reproduces the same scope of topics and time on task, but relies on customary practices of problem solving and discussion without formal debate procedures and without explicit calibration of confidence.

The pre-test is administered immediately before the start of the intervention and includes the SKM-TECH critical thinking composite, a class of probabilistic tasks with Brier score calculation, as well as brief indicators of L2 proficiency and basic statistical preparation for subsequent control. Upon completion of the instructional cycle, the post-test is administered. Written work and transcripts of oral performances are assessed by two independent raters who are blind to



participants' assignment to condition and to the measurement stage. To ensure the stability of the intervention, all tutors receive instructions, case templates and checklists for protocol adherence. A log of debate sessions is kept, recording attendance and deviations from the regulations. Threats to internal validity are reduced through cluster randomisation, standardised materials, the anonymity of participants and reviewers, and the prior alignment of assessment criteria.

Sample and context

The study involved 117 first- and second-year students of technical specialities from six academic groups, with an average of 20 students in each. English language proficiency was confirmed by an entry placement test and was not lower than B1 according to the CEFR. Groups were allocated in clusters: three to the experimental condition and three to the active control, while preserving the mathematics and ESP contact hours typical of the curriculum.

The inclusion criteria were enrolment in the target degree programmes, absence of academic debt in higher mathematics courses, consent to participate and the ability to attend classes in the established mode. In cases of absence due to air-raid alerts or blackouts, safety and learning recovery protocols were applied, and key activities were duplicated in the nearest available time slots in order to minimise systematic data loss. The learning environment corresponded [6] to the conditions of a Ukrainian university under wartime: possible interruptions, time constraints and heightened situational load, which was recorded using brief scales of cognitive load and stress.

Procedure

The intervention lasts 8 weeks in the format of 2 sessions per week and combines modules on probabilistic thinking, statistical argumentation, confidence calibration and English-language debates based on engineering cases. Measurements are carried out at four stages: pre-test, intermediate micro-measures (brief intermediate measures) during the course, post-test, delayed post-test. Written work and debate transcripts are assessed using standardised rubrics by two independent raters who are unaware of the study conditions and the measurement wave.

Intermediate micro-measures and process monitoring

- After key activities each week, a 4-item stress scale is collected for modelling the accuracy-load trade-off.
- For each written conclusion, a confidence interval and domain of use are mandatory; for oral performances, English-language markers of uncertainty are required.
- Adherence to the protocol is recorded in the session log; in the event of interruptions due to alerts, safety micro-texts and continuation markers are applied.



Delayed post-test

After 4 weeks, students are offered a short sample of equivalent tasks on statistical argumentation, calibration and one concise written debate micro-case. The main indicators are the change in the overall SKM-TECH, the dynamics of the Brier score and the stability of linguistic markers of uncertainty in L2.

Assessment instruments

The SKM-TECH critical thinking composite is a multidimensional index that aggregates four subscales: statistical argumentation, probabilistic reasoning and calibration, source verification, and argumentation in English. Tasks are scored on an analytic 0-3 scale with clear level descriptors; the subscales are normalised to 0-25 and summed to a total score of 0-100. Construct validity is tested by means of confirmatory factor analysis, and internal consistency is expected at $\alpha \geq 0.70$ for the overall composite. Form equivalence is ensured through parallel pre-test and post-test versions and prior piloting of difficulty.

Educational debate rubric (DEBATE-ENG). A scale with six 0-4 criteria for the assessment of: 1) clarity of the thesis and assumption frame, 2) evidential strength and correctness of numerical claims, 3) language of uncertainty and limits of conclusions, 4) counterarguments and refutation, 5) team interaction and moderation, 6) clarity of L2 and accuracy of terminology. The total range is 0-24. Two independent raters work in a blind mode. The rubric is aligned with the SKM-TECH composite through shared markers of evidentiality, which makes it possible to test convergent validity.

A set of 10 predictive statements is used, for each of which the participant specifies a probability in per cent. After the true outcomes have been established, the mean Brier score is calculated as

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2, \text{ where } o_i \in \{0,1\}$$

and lower values indicate better calibration. In addition, calibration curves are plotted for the intervals 0-10, 10-20, ... 90-100 and the expected calibration error (ECE) is computed. For cases with differing importance, weighted averaging can be applied.

Source micro-verification (CHECK-FAST). Four short scenarios are used to detect: 1) fake or non-resolving DOIs, 2) substitution of units of measurement, 3) manipulative graphs, 4) inconsistencies in data tables. Each scenario is scored 0-3 according to the criteria of accuracy, completeness and justification of the conclusion; the total range is 0-12. A time limit of 6-8 minutes reduces artefacts associated with searching external sources and captures the extent to which basic checks have been automatized.



Scales of cognitive load and stress. After key activities, the single-factor Paas 1-9 scale is applied for subjective cognitive load, together with a brief four-item indicator of state stress with responses on a 1-5 scale. Both scales are used as covariates in the analysis of effectiveness and for monitoring the accuracy-load trade-off. The total completion time does not exceed 30-45 seconds per wave.

Short L2 competence test. A paper-and-pencil or computer-based test of 15-20 minutes' duration combines reading with selected vocabulary and grammar tasks and is aimed at checking the B1 threshold. An overall score and level category are calculated, which are included in the models as covariates and moderators of the effect. Given a sufficient sample size, Rasch calibration may be applied to test item invariance.

Rules of aggregation and data quality. The primary indicators of effectiveness are the change in the overall SKM-TECH and the Brier score from pre-test to post-test; the secondary indicators are the SKM-TECH subscales, the DEBATE-ENG score, ECE and stress. Missing data are processed according to the principle of maximal completeness aiming to retain as many cases as possible under principled missing-data handling: the analysis plan is pre-registered, and multiple imputation or mixed-effects models are applied. All scales undergo checks of reliability and invariance between conditions, which ensures the correctness of comparisons.

Methods of analysis

Primary models. For the main indicators of effectiveness, a linear regression model of the ANCOVA type is applied:

$$Y_{post} = \beta_0 + \beta_1 Condition + \beta_2 Y_{pre} + \beta_3 L2 + \beta_4 Instructor + \varepsilon,$$

where Y_{post} is the outcome after the experiment or intervention;

$Condition$ is the experimental condition (for example, control or experimental group);

Y_{pre} is the value of the indicator before the intervention (a covariate that controls for initial differences);

$L2$ is a possible variable that indicates, for example, second language proficiency (a binary or categorical variable);

$Instructor$ is a variable that takes into account the effect of different instructors (a factor);

ε is the random error term; $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are the regression parameters that indicate the contribution of each variable included in the equation.

The model estimates how the experimental condition ($Condition$) affects the final outcome while controlling for the influence of the initial level, second

language proficiency and instructor. It is necessary to test whether Condition has a significant effect on the outcome after controlling for other factors. In this context, the key parameter is β_1 . If it is statistically significant ($p < 0,05$), this means that the condition has an effect even after controlling for the other variables.

Multilevel models.

For outcomes with repeated measurements and a hierarchical data structure (students nested within groups), linear mixed models with random intercepts for groups are estimated:

$$Y_{it} = \gamma_0 + \gamma_1 \text{Condition}_i + \gamma_2 \text{Time}_t + \gamma_3 Y_{pre,i} + \gamma' X_i + u_{group} + e_{it},$$

where Y_{it} is the dependent variable (the outcome for individual i at time t);

γ_0 is the constant (intercept), that is, the baseline level when all variables = 0 (fixed effect);

Condition_i – is the experimental or control condition (independent variable);

Time_t is the time point (pre, post, different waves of the study) (independent variable);

$Y_{pre,i}$ is the previous value of the variable (baseline level) (control variable);

X_i is the vector of additional explanatory variables (age, gender, education etc.) (independent variables);

γ' is the vector of coefficients for X_i (fixed effects);

u_{group} is the random effect of the group (class, company, country etc.).

This model combines fixed and random effects. The fixed effects (γ) show the average impact of the factors on the outcomes for the entire sample, while the random effects u_{group} take into account specific differences that are not explained by the measured variables. In other words, each group has its own regression line with a small shift.

The model estimates the relationship between the outcome and several factors, taking into account that some factors exert a systematic influence (fixed effects), others vary randomly from group to group (random effects), and there are also individual fluctuations.

Diagnostics include residual analysis, checks of linearity, homoscedasticity and influential points.

For the between-condition difference after adjustment for the pre-test, *Hedges' g* is computed as an adjusted estimate of the difference between groups with the small-sample correction J :



$$g = J \cdot \frac{\bar{Y}_{post,E} - \bar{Y}_{post,C}}{SD_{pooled,res}},$$

where $\bar{Y}_{post,E}$ is the mean value of the dependent variable in the experimental group at post-test;

$\bar{Y}_{post,C}$ is the mean value in the control group at post-test;

$SD_{pooled,res}$ is the pooled standard deviation for the two groups, that is, the average variability of the outcomes;

J is the correction coefficient that compensates for bias in small samples.

Hedges' g shows how strongly the mean values of the two groups differ:

if $g = 0$ there is no difference between the groups;

if $g = 0,5$ the experimental group is better by half a standard deviation;

if $g = 1$ the effect is strong.

The internal consistency of the SKM-TECH composite and subscales is evaluated using Cronbach's α with 95% confidence intervals for α . The target threshold is $\alpha \geq 0.70$ for the summary indices. If lower values are obtained, an item contribution analysis is carried out and the descriptors are reviewed.

For the set of predictive tasks, calibration curves are additionally assessed for the intervals 0-10-...-90-100 and the Expected Calibration Error. Comparisons between conditions are performed within the framework of mixed models with random intercepts for groups, or via ANCOVA for averaged individual indicators.

Results

After adjustment for the pre-test in ANCOVA, the experimental condition showed higher outcomes compared to the active control, both on the overall SKM-TECH composite and on all subscales. The largest gains were observed in the subscales "probabilistic reasoning and calibration" and "source verification", where the effects reached at least a medium size according to Hedges' g . In "statistical argumentation" stable, statistically significant gains of small-to-medium magnitude were recorded. The "argumentation in English" subscale also improved, although the effects were more modest and varied across groups, which is consistent with the increased linguistic load in L2. The assessment of written and oral work showed satisfactory internal consistency of the composite and high inter-rater agreement on the DEBATE-ENG rubric.

For the predictive tasks, the Brier score in the experimental condition decreased from pre-test to post-test more substantially than in the control. The calibration curves moved closer to the bisector in the 40-60 and 60-80 intervals, which was accompanied by a reduction in the Expected Calibration Error.



Residual errors were most frequently associated with overconfidence in the high 80-100 bins; however, even in this range an improvement relative to the control was observed. Sensitivity analysis with cluster-robust standard errors at group level confirmed the robustness of these conclusions.

Models with the Condition $\times$ Time interaction showed a favourable trajectory for the experimental condition: increases in evidentiality and decreases in the Brier score occurred under unchanged or lower load at the final stages. This is consistent with the hypothesis of a learned transition from formal computations to concise, structured communication of arguments.

It should be noted that the effects persisted after controlling for L2 in ANCOVA models and multilevel specifications with random intercepts for groups. The Condition $\times$ L2 interaction for the overall SKM-TECH composite was not systematic, which indicates a general benefit of the intervention for students with different profiles of English language proficiency, provided a threshold level of B1 is met. At the same time, a moderate moderation effect was found for the "argumentation in English" subscale: students with higher L2 levels more quickly achieved clear language of uncertainty and stronger refutation, whereas students at level B1 showed [7] larger gains in confidence calibration and source checking, compensating for linguistic limitations through formalised protocols. Taken together, the results confirm that the integration of probability theory with debates strengthens both the evidential strength of reasoning and the linguistic transparency of argumentation without a substantial "cost" in the form of additional load.

The results obtained indicate that the combination of probability theory with educational debates creates a robust loop of "calculation – interpretation – communication – refutation – revision of beliefs". The largest gains in probabilistic reasoning and source verification are consistent with a mechanism in which Bayesian updating and work with confidence intervals provide the formal basis, while debates force numerical results to be translated into public, transparent arguments with obligatory delineation of the limits of conclusions. The reduction in the Brier score without a substantial increase in subjective load points to the gradual automatization of key operations and successful management of extraneous load through templates for the language of uncertainty, collocation cards and formal procedures.

More moderate effects on the English-language argumentation subscale are to be expected given the bilingual cognitive cost. At the same time, moderation by L2 shows that students with a threshold level of B1 benefit primarily through confidence calibration and the discipline of source checking, whereas higher L2 levels more rapidly convert statistical conclusions into precise and concise English formulations of boundaries.



This supports the hypothesis of complementary pathways: formal rigour reduces the space for logical errors, while structured communication reduces the risk of misinterpretation under time pressure.

A key contextual observation concerns the role of English-language argumentation under the stress of air-raid alerts. The implemented safety micro-texts and continuation markers made it possible to preserve the validity of assessment and the rhythm of discussion after interruptions. During the first debates, load increased, but in the second cycle it normalised, which is consistent with a training effect: students internalise the procedures, eliminate superfluous operations and move towards an "economical" language of evidence. Thus, English-language argumentation in a formalised format proves not to be a weak link but an instrument for stabilising thinking under stress, provided linguistic templates and clear roles are ensured.

In courses on power engineering and systems reliability, debates can be replaced by design or safety reviews with the same requirements regarding the limits of conclusions and sensitivity analysis. In cybersecurity, the format is easily adapted to incident analysis and the comparison of response policies, where Bayesian updating is combined with risk matrices. In software engineering, the role of debates can be played by structured code reviews or architecture defences with validation of assumptions and performance metrics. The general principle remains unchanged: numerical evidentiality is first constructed, then communicated, then subjected to refutation and only thereafter translated into decisions. The results obtained confirm that the integration of probability theory with debates creates [8] a stable mechanism for the development of critical thinking in a wartime context.

The model reduces typical cognitive biases, increases the transparency of argumentation and does not impose an excessive "cost" in the form of additional load, provided that principles of load management are observed and language support is ensured. This opens up scope for scaling to adjacent technical disciplines and for building interdisciplinary tracks in which students learn to think and speak in the language of probability theory as confidently as in the language of formulae. In this study, randomisation was carried out at group level, so differences between clusters in motivation, timetable, composition of specialities and contact hours are possible. This creates a risk of biased estimates of effectiveness. Mitigation involves the use of cluster-robust standard errors, multilevel models with random intercepts for groups, the inclusion of fixed effects for timetable and programme, and sensitivity analysis with the removal of individual high-influence clusters.

At the same time, variations in teaching style, experience with debates and adherence to the protocol may partly explain the observed differences. For control

purposes, unified materials, pre-start training, protocol adherence checklists and blind review of work were applied. In the analytical models, instructor was included as a covariate or as a level of the hierarchy. Nevertheless, residual instructor effects cannot be completely eliminated, which limits internal validity.

Baseline preparation in probability theory and mathematical statistics affects the speed of mastering Bayesian transformations, confidence intervals and hypothesis tests. Pre-test indicators of statistical literacy are included as a covariate, and stratified checks of effects are carried out in subgroups with different baseline levels. Taken together, these factors do not nullify the observed effects but define the boundaries for their cautious interpretation. The proposed statistical corrections and procedural safeguards reduce bias, but multi-site replications, preregistration of the protocol and independent monitoring of compliance with the methodology are advisable to strengthen the validity of the results.

Conclusions and practical recommendations

Conclusions. The integration of probability theory with educational debates in English forms [9] a stable loop of justification: from calculations and interpretation to public communication, refutation and revision of beliefs. The empirical effects manifest themselves in gains on the critical thinking composite and subscales, improved calibration of probabilistic estimates and reduced reliance on rhetorical devices without an increase in subjective load. Under wartime interruptions to learning, structured English-language argumentation functions as a mechanism for stabilising thinking and preserving the validity of assessment.

Table 1

Implementation guide for the higher mathematics and ESP course

Component	Main focus	Key tools / artefacts (approx. count)	Primary focus sessions (out of 16)	Suggested weight in course grade (%)
Organisational co-teaching model	Coordination between mathematics and ESP instructors; shared planning	2 instructors; 1 joint syllabus; 1 weekly coordination meeting	2	5
Higher mathematics component	Formal probabilistic methods, confidence intervals, hypothesis testing, risk	4 core tool sets: frequency cards; odds-ratio worksheets; interval calculator; graph- and-units checklist	6	35



Component	Main focus	Key tools / artefacts (approx. count)	Primary focus sessions (out of 16)	Suggested weight in course grade (%)
ESP component	L2 argumentation, framing uncertainty, debate language and structure	5+ language supports: L2 templates; argument maps; debate scripts; sample micro-texts; collocation lists	4	25
Assessment and feedback	Measurement of SKM-TECH, calibration tasks, debate performance feedback	3 main instruments: SKM-TECH battery; 1 prediction set with Brier score; DEBATE-ENG rubric	2	20
Load management and linguistic support	Monitoring and reducing extraneous load; scaffolding for level B1 students	4 support tools: Paas form; 4-item stress scale; worked-example set; B1 collocation cards	1	10
Quality assurance and reproducibility	Protocol fidelity, standardisation, robustness checks	3 QA tools: instructor training pack; protocol checklist and Fidelity Log; model specification templates	1	5

Source: created by authors

Table 2

Safety instructions and micro-texts in case of interruptions to classes

Phase	Role	Micro-text / core instruction	Key actions / notes
Pause for an alert	Teacher	Attention, air-raid alert. We stop the discussion. We record STOP-MARK: team, speaker, thesis, time. We proceed in an orderly manner to the shelter.	Stop discussion; record STOP-MARK; move to shelter in an orderly way.
In the shelter	Moderator	CONT-MARK: next to speak is team B, speaker 2, topic – limits of the conclusion and sensitivity analysis. Write down one counterargument and one numerical refinement each.	Maintain discussion structure; assign next speaker; ensure each student writes one counterargument and one numerical refinement.

Phase	Role	Micro-text / core instruction	Key actions / notes
Return to the classroom	Teacher	We resume from CONT-MARK. We have 2 minutes to reconstruct the arguments. We continue with team B, speaker 2. The time regulations remain unchanged.	Reconstruct arguments in 2 minutes; resume turn-taking according to CONT-MARK; keep original time limits.
Preserving the validity of assessment	Teacher / raters	Before the break, record what has been assessed and what remains. After return, assess only new contributions. If the session is not completed, transfer the remaining part to the nearest available slot, labelling the continuation as "Debate 1b" or "Lab 2b".	Document assessment status pre-break; avoid double marking; clearly label continuation sessions.

Source: created by authors

Minimal starter package for the course.

1. A syllabus with integrated goals, a set of cases with explicit base rates and data.
2. Frequency and probabilistic language cards, worksheets for confidence intervals and Bayesian transformations.
3. SKM-TECH and DEBATE-ENG rubrics, transcript templates and argument maps.
4. Forms for Paas and stress, STOP-MARK and CONT-MARK protocols, a fidelity log.

Scaling and support.

1. Start with a single 4-week module, followed by expansion to a full 8-week cycle.
2. Hold weekly methodological reviews of recordings and artefacts with alignment of criteria.
3. Summarise results in internal departmental reports tracking key indicators: SKM-TECH, Brier score, Paas, ICC.

The proposed guide ensures coherent implementation of the model in higher mathematics and ESP courses, supports safety and continuity of learning during interruptions, and creates conditions for reproducible and valid measurements of the development of critical thinking.

In summary, the study demonstrates that integrating probabilistic modelling with structured debates leads to statistically significant improvements in both the accuracy and communicative clarity of students' reasoning.



The intervention strengthens the link between numerical evidence and transparent verbal justifications, thereby reducing typical cognitive biases and overconfidence in high-stakes judgments. Overall, the results suggest that this combined approach can serve as a scalable template for critical thinking instruction in technical education under conditions of chronic stress and disruption.

References

1. Khomyuk, I. V., Sachaniuk-Kavetska, N.V., Kyrylashchuk, S.A., & Khomyuk, V.V. (2024). Formuvannia lohiko-analitychnoi kompetentnosti maibutnikh inzheneriv na zaniattiakh z vyshchoi matematyky [Formation of logical-analytical competence of future engineers in higher mathematics classes]. *Aktualni pytannia pryrodnycho-matematychnoi osvity: zbirnyk naukovykh prats*, 1(23), 104-111. <https://doi.org/10.5281/zenodo.12160425>. [in Ukrainian].
2. Sachaniuk-Kavetska, N. V., & Kavetskyi, V. V. (2021). Zastosuvannia kryteriiu Fishera dlia zabezpechennia dostovirnosti rezultativ otsiniuvannia zalyshkovykh znan studentiv. [Application of Fisher's criterion to ensure the reliability of the results of assessing students' residual knowledge]. *Fyzyko-matematychna osvita*, 28(2), 71-76. <https://doi.org/10.31110/2413-1571-2021-028-2-012>. [in Ukrainian].
3. Stepanova, I. S., Nykyporets, S. S., & Kukharchuk, H. V. (2025). Integrating artificial intelligence tools into project-based English language instruction for technical students: A framework for fostering critical and creative thinking. In A. Ostenda & D. Kalita (Eds.), *Innovation-driven development in education, digital economy, and applied technologies* (pp. 208-215). The University of Technology in Katowice Press. <https://doi.org/10.54264/M055>.
4. Nykyporets, S. S., Kot, S. O., Hadaichuk, N. M., Herasymenko, N. V., & Kukharchuk, H. V. (2025). Advancing critical thinking skills in future engineers through the analysis of AI-generated and AI-verified English-language sources. *Current Issues in Modern Science. Series "Pedagogy"*, 8(38), 1143-1155. [https://doi.org/10.52058/2786-6300-2025-8\(38\)-1143-1155](https://doi.org/10.52058/2786-6300-2025-8(38)-1143-1155).
5. Rainsford, E., & Foster, C. (2025). Confidence assessment, calibration, and gender. *International Journal of Science and Mathematics Education*. <https://doi.org/10.1007/s10763-025-10615-1>.
6. Nykyporets, S.S., Melnyk, O.D., Ibrahimova, L.V., Hadaichuk, N.M., & Derun, V.H. (2024). Advancing critical thinking skills among higher education students through English language instruction: Contemporary approaches and strategies. *Prospects and Innovations of Science. Series "Pedagogy"*, 1(35), 34-45. [https://doi.org/10.52058/2786-4952-2024-1\(35\)-34-45](https://doi.org/10.52058/2786-4952-2024-1(35)-34-45).
7. Nykyporets, S.S., Melnyk, O.D., Ibrahimova, L.V., Boiko, Y.V., & Kukharchuk, H.V. (2023). Fostering critical thinking in technical university students in foreign language classes: Strategies and approaches for cultivating analytical proficiency. *Bulletin of Science and Education. Series "Pedagogy"*, 8(14), 344-360. [https://doi.org/10.52058/2786-6165-2023-8\(14\)-344-360](https://doi.org/10.52058/2786-6165-2023-8(14)-344-360).
8. Friedman, H. H. (2023). Cognitive Biases and Their Influence on Critical Thinking and Scientific Reasoning: A Practical Guide for Students and Teachers [Preprint]. SSRN. <http://dx.doi.org/10.2139/ssrn.2958800>.
9. Betz G, Richardson K (2023). Probabilistic coherence, logical consistency, and Bayesian learning: Neural language models as epistemic agents. *PLoS ONE* 18(2): e0281372. <https://doi.org/10.1371/journal.pone.0281372>.