



## OPEN ACCESS

## EDITED BY

Nivaldo A. P. de Vasconcelos,  
Federal University of Pernambuco, Brazil

## REVIEWED BY

David Escobar-Castillejos,  
Panamerican University, Mexico  
Mert Başaranoğlu,  
Mersin University, Türkiye

## \*CORRESPONDENCE

Nuno Silva Gonçalves  
✉ nunogoncalves@med.uminho.pt

RECEIVED 29 October 2025

REVISED 24 November 2025

ACCEPTED 26 November 2025

PUBLISHED 06 January 2026

## CITATION

Silva Gonçalves N, Collares C and  
Pêgo JM (2026) AI-enhanced adaptive testing  
with cognitive diagnostic feedback and its  
association with performance in  
undergraduate surgical education: a pilot  
study.

*Front. Behav. Neurosci.* 19:1735237.

doi: 10.3389/fnbeh.2025.1735237

## COPYRIGHT

© 2026 Silva Gonçalves, Collares and Pêgo.  
This is an open-access article distributed  
under the terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other forums is  
permitted, provided the original author(s) and  
the copyright owner(s) are credited and that  
the original publication in this journal is cited,  
in accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# AI-enhanced adaptive testing with cognitive diagnostic feedback and its association with performance in undergraduate surgical education: a pilot study

Nuno Silva Gonçalves<sup>1,2\*</sup>, Carlos Collares<sup>1,3,4,5,6</sup> and  
José Miguel Pêgo<sup>1,2,4,7</sup>

<sup>1</sup>Life and Health Sciences Research Institute (ICVS), University of Minho, Braga, Portugal, <sup>2</sup>ICVS/3B's, PT Government Associate Laboratory, Braga, Portugal, <sup>3</sup>European Board of Medical Assessors, Cardiff, United Kingdom, <sup>4</sup>Inspirali Educação, São Paulo, Brazil, <sup>5</sup>Faculdades Pequeno Príncipe, Curitiba, Brazil, <sup>6</sup>Medical Education Unit, Faculty of Medicine and Biomedical Sciences, University of Algarve, Faro, Portugal, <sup>7</sup>iCognitus4ALL – IT Solutions, Porto, Portugal

**Background:** Effective feedback in the cognitive domain is essential for surgical education but often limited by resource constraints and traditional assessment formats. Artificial Intelligence (AI) has emerged as a catalyst for innovation, enabling automated feedback, real-time cognitive diagnostics, and scalable item generation, thereby transforming how future surgeons learn and are assessed.

**Methods:** An item bank of 150 multiple-choice questions was developed using AI-assisted item generation and difficulty estimation. A formative Computerized Adaptive Testing (CAT), balanced across three cognitive domains (memory, analysis, and decision) and surgical topics, was delivered via QuizOne® 3–5 days before the summative Progress Test. A total of 147 students participated, of whom 116 completed the formative CAT. Performance correlations, group comparisons, analysis of covariance (ANCOVA), and regression analyses were conducted.

**Results:** Students who voluntarily completed CAT showed higher Progress Test scores, though causality cannot be established due to self-selection bias ( $p = 0.021$ ), with the effect persisting after adjusting for prior academic performance (ANCOVA  $p = 0.041$ ). Memory skills were the strongest predictors of summative outcomes ( $R^2 = 0.180$ ,  $\beta = 0.425$ ), followed by analysis ( $R^2 = 0.080$ ,  $\beta = 0.283$ ); decision was not significant ( $R^2 = 0.029$ ,  $\beta = 0.170$ ).

**Conclusion:** AI-enhanced CAT–Cognitive Diagnostic Modeling (CDM) represents a promising formative approach in undergraduate surgical education, being associated with higher summative performance and providing individualized diagnostic feedback. Refining feedback presentation and enhancing decision-making assessment could further optimize its educational impact.

## KEYWORDS

artificial intelligence, computerized adaptive testing, cognitive diagnostic modeling, surgical education, feedback, cognitive skills, assessment innovation, educational technology

## Introduction

Artificial Intelligence (AI) is transforming medical education by enabling automated item generation, difficulty estimation, and individualized feedback that was previously impractical (Pohn et al., 2025; Shaw et al., 2025; Gordon et al., 2024; Mir et al., 2023). When integrated with adaptive testing and cognitive diagnostic models, AI allows feedback to evolve from static score reporting into a continuous, formative process that guides learning (Sunmboye et al., 2025). This approach is particularly valuable in surgical education, where cognitive, technical, and non-technical skills intersect and are required, allowing educators to identify specific strengths and weaknesses and promote deliberate practice toward surgical competence (Gomez et al., 2025; Rosendal et al., 2023; Ounounou et al., 2019; Dedy et al., 2013).

Although feedback is widely recognized as a key driver of learning, its application in the cognitive domain, particularly during undergraduate surgical education, remains underexplored (Garner et al., 2014; El Boghdady and Alijani, 2017). Traditional mechanisms often fail to provide timely, specific, and actionable insights into students' cognitive performance, especially in complex domains such as reasoning and decision-making (Burgess et al., 2020; Shaughness et al., 2017). As medical curricula increasingly adopt competency-based models, adopting assessment strategies that deliver targeted, data-informed feedback has become essential to enhance learning and self-regulation (Ross et al., 2022).

Cognitive competence involves integrating basic scientific knowledge with clinical information to interpret findings and make informed decisions under uncertainty, skills that are critical in surgical practice (Crebbin et al., 2013; Madani et al., 2017). However, traditional multiple-choice exams often emphasize factual recall, provide delayed feedback, and fail to capture deeper levels of reasoning (Butler and Roediger, 2008).

Providing feedback in summative assessments remains a challenge in medical education. Concerns over item security, fairness, and resource limitations often lead institutions to restrict item-level feedback (Appelhaus et al., 2023; Harrison, 2017). Consequently, summative assessments frequently become "black boxes," offering scores without meaningful guidance and reinforcing a culture of performance rather than development.

When integrated into assessment systems, AI can support the creation of high-quality questions, estimate item difficulty, and generate individualized diagnostic feedback with minimal faculty effort. These capabilities complement Computerized Adaptive Testing (CAT), a psychometric method proposed by Lord (1971), Owen (1975), and Chang and Ying (1996) that dynamically adjusts item difficulty based on student responses (Chang and Ying, 1996). By tailoring the test to the learner's ability level, CAT increases the efficiency and precision of assessment with fewer items, reducing test fatigue while maximizing information. It can also be useful in determining the true score and competence of an examinee (Collares and Cecilio-Fernandes, 2019; Burr et al., 2016). In the context of medical education, CAT has been successfully implemented in progress testing, licensing exams, and residency selection processes (Burr et al., 2016; Xu et al., 2023; Van Wijk et al., n.d.; Seo et al., 2024). Its potential as a formative tool for learning, however, remains underutilized.

CAT can also be coupled with Cognitive Diagnostic Modeling (CDM), a psychometric framework that analyzes student responses to infer the proficiency of specific cognitive attributes. Unlike classical test theory, which provides a single overall score, CDM enables a multidimensional understanding of performance by categorizing items and responses according to specific cognitive processes. For instance, an item might assess recall of factual knowledge, interpretation of clinical signs, or the application of pathophysiological reasoning. By classifying and analyzing items in this way, CDM supports the generation of individualized feedback, offering students a roadmap for targeted improvement (Ma et al., 2023; Williamson, n.d.; Leighton and Gierl, 2007; Barthakur et al., 2022). This granular diagnostic capability is particularly relevant in complex curricular areas like surgery, where different cognitive skills are needed to approach diverse clinical scenarios. In this study, items were categorized into three cognitive diagnostic models (memory, analysis, and decision) reflecting the cognitive tasks proposed by the National Board of Medical Examiners (NBME) for assessing medical knowledge application (Billings et al., n.d.). These domains were selected because they align with the NBME's framework for evaluating progressively complex levels of cognitive processing, ranging from factual recall to clinical reasoning and decision-making, which are particularly relevant in the context of surgical education.

Despite their theoretical advantages, CAT and CDM have rarely been explored as learning tools in undergraduate surgical education. By evaluating an AI-supported CAT-CDM intervention in a real educational setting, this study aims to advance data-informed feedback practices, improve alignment between formative and summative assessment, and support the development of self-regulated surgical learners.

## Materials and methods

### Item development and cognitive models

An item bank was developed using Automated Item Generation (AIGen) techniques, guided by cognitive models aligned with the exam blueprint (Leighton and Gierl, 2007; Falcão et al., 2023; Gierl et al., 2022; Falcão et al., 2022). The overall workflow of AI-assisted item generation, expert validation, and calibration is illustrated in Figure 1. Each item was also categorized according to three cognitive diagnostic models based on the NBME framework (Billings et al., n.d.): memory (questions that require recall of factual information), analysis (questions that demand interpretation of clinical information), or decision (questions that require synthesis and clinical critical decision). To ensure consistency, two independent raters with experience in medical education and assessment independently classified all items according to these categories. Inter-rater reliability for initial classification was substantial (Cohen's  $\kappa = 0.76$ ), indicating good agreement. Discrepancies were resolved through discussion until consensus was achieved.

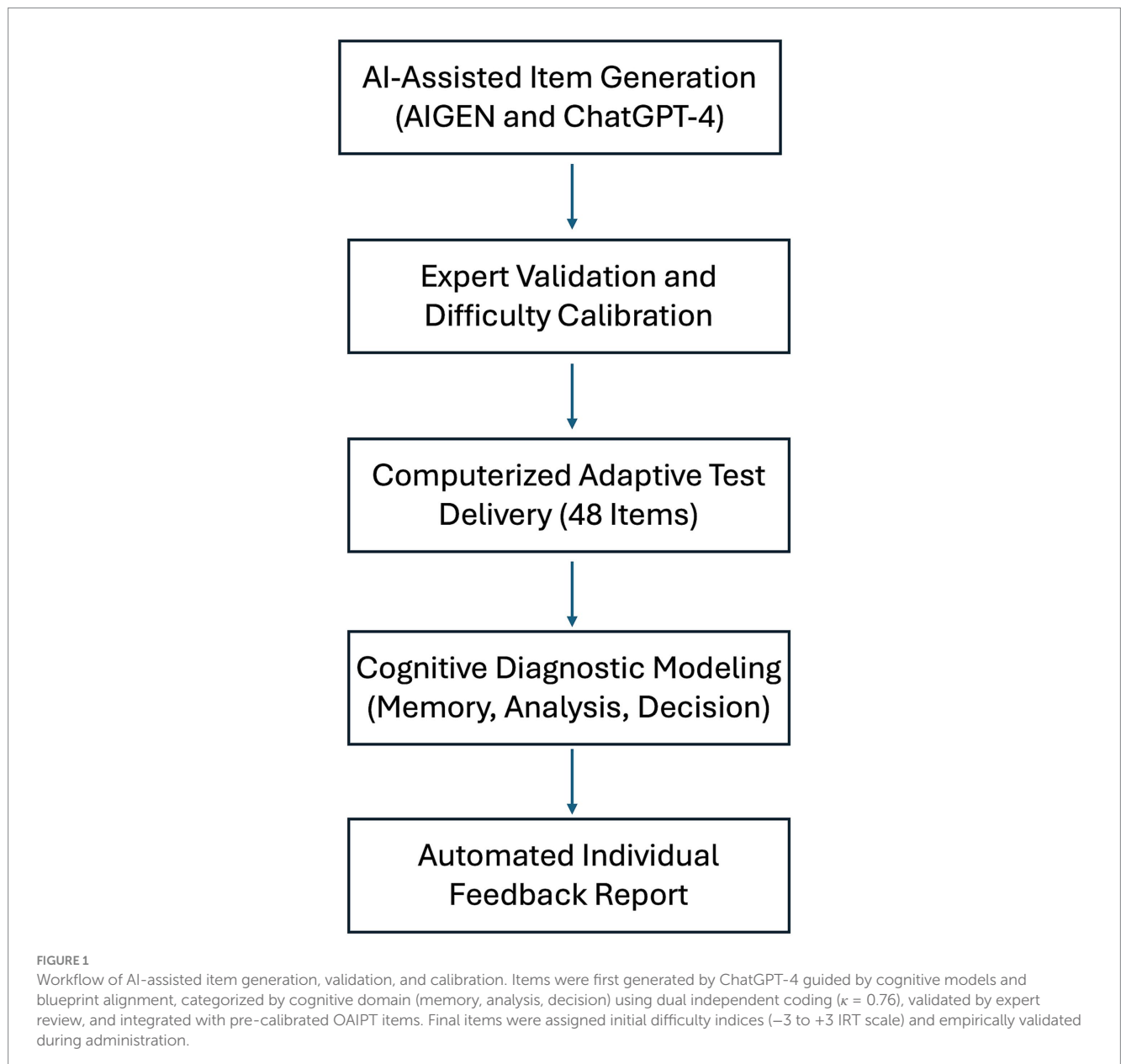
Additionally, a retrospective double-blind classification of a random sample of 25 items by two external raters who were not involved in the original coding. Inter-rater reliability for this retrospective validation was Cohen's  $\kappa = 0.68$ , indicating substantial agreement and supporting the consistency of domain assignment. In

addition to newly generated items, calibrated items from the existing bank of the Online Adaptive International Progress Test (OAIPT) were incorporated, taking advantage of their known difficulty parameters and item quality. For the newly created items, difficulty indices (ranging from  $-3$  to  $+3$  on the IRT scale) were initially estimated using ChatGPT-4 (OpenAI, web interface) with the following structured prompt: “Given the previous examples with their respective difficulty indices, estimate the expected difficulty of the following multiple-choice question for 5th-year medical students in a summative exam.” Each AI-generated estimate and rationale were reviewed by two content experts for accuracy and consistency. Items were accepted only when both experts agreed that (a) the content was clinically correct, (b) the reasoning aligned with the intended cognitive domain, and (c) the predicted difficulty was plausible compared with similar calibrated items. Disagreements were resolved through discussion. This process produced an initial difficulty

calibration for new items before empirical validation during test administration. To enhance transparency, representative examples of each cognitive domain are provided in [Supplementary material 1](#), allowing readers to evaluate the appropriateness of domain assignment.

## Platform setup

The digital assessment platform QuizOne® was used to deliver CAT via its dedicated module (Rice et al., 2022). This platform enabled dynamic test adaptation based on individual student responses. Weighted Likelihood (WL) method was used as the theta estimator, providing robust estimates of student ability even with a small number of items. For the selection of subsequent items, the platform applied the Maximum Fisher Information (MFI) criterion, which chooses the



next item that maximizes the expected information gain at the student's ability level (Chang and Ying, 1996).

## Formative CAT exam design

A formative Computerized Adaptive Test (CAT) consisting of 48 multiple-choice questions (MCQs) was created for this study and made available to students 3–5 days before the summative Progress Test of the Surgical Curricular Unit. The CAT was delivered through the platform and integrated Cognitive Diagnostic Modeling (CDM) to classify each item by cognitive domain—memory, analysis, or decision. The platform was programmed to present one item from each cognitive domain across 16 surgical topics defined in the exam blueprint: Urology, Trauma, Abdominal Wall, Orthopedics, Otorhinolaryngology, Ophthalmology, Neurosurgery, Hepatobiliopancreatic, Esophagogastric, Endocrine and Breast, Colorectal, and Vascular Surgery. Each student therefore, completed exactly 48 items, as the adaptive algorithm was configured with a fixed-length stopping rule to ensure comparable duration and psychometric precision across participants. The formative CAT aimed to reinforce learning by identifying specific cognitive gaps and providing individualized feedback on students' performance profiles. While there was curricular overlap in learning objectives, the specific items used in the formative CAT were distinct from those included in the summative exam, preventing direct content duplication.

Additionally, the authors collected the reports of the previous two exams of the surgical curricular unit (Exam 1 and Exam 2) done by the same student sample.

## Participants

All the students who enrolled in the surgical curricular unit were invited to participate in the formative exam. A convenience sample of 116 volunteer students completed the formative CAT exam.

## Data collection and analysis

Item-level performance data and overall test scores were extracted from the platform. Statistical analysis was performed using JASP (version 0.18.1).

Before inferential testing, the distribution of all continuous variables was assessed using the Shapiro–Wilk test to evaluate assumptions of normality. Normality was tested with the Shapiro–Wilk test. When assumptions of normality were violated, non-parametric tests were applied (Spearman's  $\rho$  for correlations, Mann–Whitney  $U$  for between-group comparisons, Kruskal–Wallis  $H$  for multiple groups). Linear regression assumptions were examined by inspecting residual plots and variance inflation factors. All analyses were independently verified by two authors to ensure consistency and accuracy of results. To examine relationships between cognitive performance indicators and summative outcomes, we conducted correlation analyses between final CAT scores, CDM sub-scores, and summative exam grades. Group comparisons were performed to assess differences between students who completed the formative CAT exam and those who did not. Additionally, ANCOVA models were

used to evaluate the independent effect of CAT participation on summative performance, adjusting for prior academic performance. Finally, simple linear regressions were performed to identify predictors of summative exam scores, including the contribution of specific CDM sub-scores and overall test performance. Statistical significance was set at  $p < 0.05$ .

## Feedback delivery

Upon completion of the formative CAT, each student received an individualized automated feedback report generated by the platform. The report began with a brief explanation of the standardized scoring system (mean = 500, SD = 100) and an illustrative graph clarifying how to interpret scores (e.g., 600 = 1 SD above average). Subsequent pages presented tables and bar charts displaying the student's standardized scores across multiple dimensions: 1) System categories (e.g., respiratory, digestive, musculoskeletal); 2) Surgical disciplines (e.g., hepatobiliopancreatic, colorectal, trauma, vascular); 3) Medical competencies (diagnosis, treatment, scientific principles, management, pathophysiology); 4) Cognitive domains (memory, analysis, decision).

## Satisfaction questionnaire

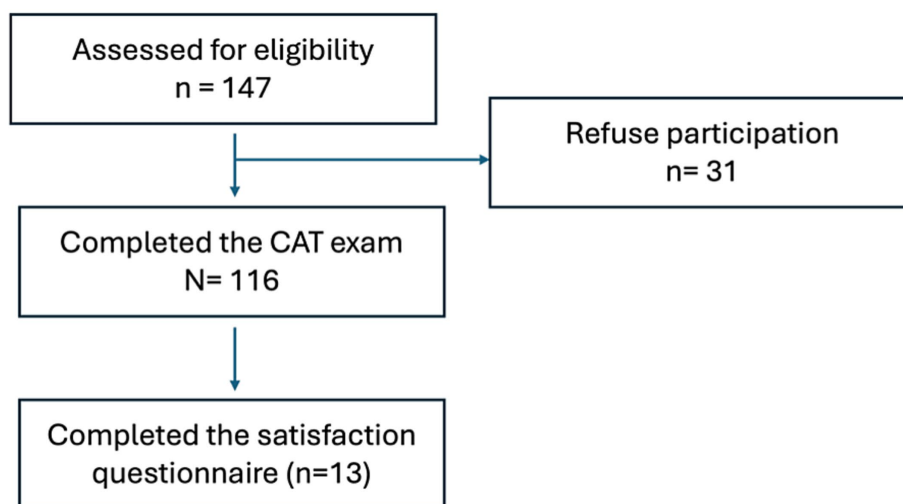
At the end of the curricular unit, participating students completed a satisfaction questionnaire assessing the usability, perceived usefulness, and impact of CAT and CDM on their learning process. The questionnaire consisted of nine Likert-scale items (1 = least positive; 5 = most positive) and two open-ended questions addressing the most positive and least positive aspects of the test and feedback. However, the response rate was insufficient (13%, 15 of 116 participants) to allow meaningful analysis; therefore, these data were not included in the present report.

## Ethical commission approval

This study was conducted following the ethical principles outlined in the Declaration of Helsinki. Ethical approval was obtained from the Ethics Committee for Research in Life and Health Sciences, University of Minho. Participation was voluntary, and all students provided informed consent before inclusion in the study. No identifiable personal data was collected. Participants were informed that their academic evaluation would not be influenced by their decision to participate or by their performance in the formative assessments. Data were stored securely and used exclusively for research purposes.

## Results

The final item bank used in the CAT exam was composed of 150 MCQ items, 84 extracted from the previously calibrated database (OAIPT) and 66 created by faculty members, experienced in item writing. From that item bank, the software developed an individual customized exam with 48 questions, tailored to the student's level of competence, with a matching blueprint of the same topics as the



**FIGURE 2**  
CONSORT-style flowchart illustrating participant inclusion and attrition. Of the 147 students initially enrolled in the Surgical Curricular Unit, 116 voluntarily completed the formative Computerized Adaptive Test (CAT). Thirty-one students did not participate. Among the CAT participants, 15 completed the post-course satisfaction survey.

**TABLE 1** Descriptive statistics for key variables, including mean, standard deviation, and observed range values for the formative CAT total score, cognitive domain sub-scores, and summative assessments.

Variable	Mean	SD	Minimum	Maximum
Final CAT score	589.95	46.44	447.00	698.00
Skill–Memory	604.79	54.87	446.00	713.00
Skill–Analysis	564.94	57.71	434.00	694.00
Skill–Decision	551.11	33.05	436.00	671.00
Progress Test score	14.91	2.16	8.80	18.20
Exam 1 score	12.98	2.07	5.70	17.10
Exam 2 score	12.49	2.47	6.20	17.20

progress test and the previous surgical curricular unit. The 48 questions were equally divided between the three different diagnostic models: decision, memory, and analysis.

A total of 147 students were included in the analysis, of whom 116 completed the formative CAT. The participant inclusion process and attrition are illustrated in Figure 2. A post-hoc power analysis for the ANCOVA comparing CAT participants ( $n = 116$ ) and non-participants ( $n = 31$ ), with  $\alpha = 0.05$ ,  $f = 0.173$  (derived from  $\eta^2 = 0.029$ ) and two covariates, indicated approximately 0.49 power to detect this effect size. This suggests limited sensitivity for small effects, consistent with the exploratory nature of the study.

Normality was evaluated using the Shapiro–Wilk test. Most variables, including Progress Test, Exam 1, and Exam 2, and CDM-specific measures, deviated from normality in at least one group ( $p < 0.05$ ). Therefore, non-parametric tests were applied when assumptions were not met. As expected, grades from the CAT exam followed a normal distribution. Table 1 summarizes the descriptive

statistics for all key variables, including the overall CAT score, cognitive sub-scores, and summative exam results.

Spearman’s rank correlations were computed to examine the relationships between the final CAT score, cognitive diagnostic sub-scores (Skill–Memory, Skill–Analysis, Skill–Decision), and academic outcomes (Exam 1, Exam 2, and Progress Test). Results are detailed in Table 2 and visually represented in Figure 3.

The final CAT score showed strong correlations with Skill–Memory ( $\rho = 0.73$ ,  $p < 0.001$ ) and Skill–Analysis ( $\rho = 0.63$ ,  $p < 0.001$ ), and a moderate correlation with Skill–Decision ( $\rho = 0.42$ ,  $p < 0.001$ ).

Regarding academic outcomes, the Progress Test was moderately correlated with Skill–Memory ( $\rho = 0.38$ ,  $p < 0.001$ ) and showed weaker associations with Skill–Analysis ( $\rho = 0.26$ ,  $p = 0.006$ ) and Skill–Decision ( $\rho = 0.28$ ,  $p = 0.003$ ). Both Exam 1 and Exam 2 correlated more strongly with Skill–Memory than with the other sub-scores. Additionally, Exam 1, Exam 2, and the Progress Test were highly intercorrelated ( $\rho \approx 0.8$ ,  $p < 0.001$ ), indicating consistent performance across summative assessments.

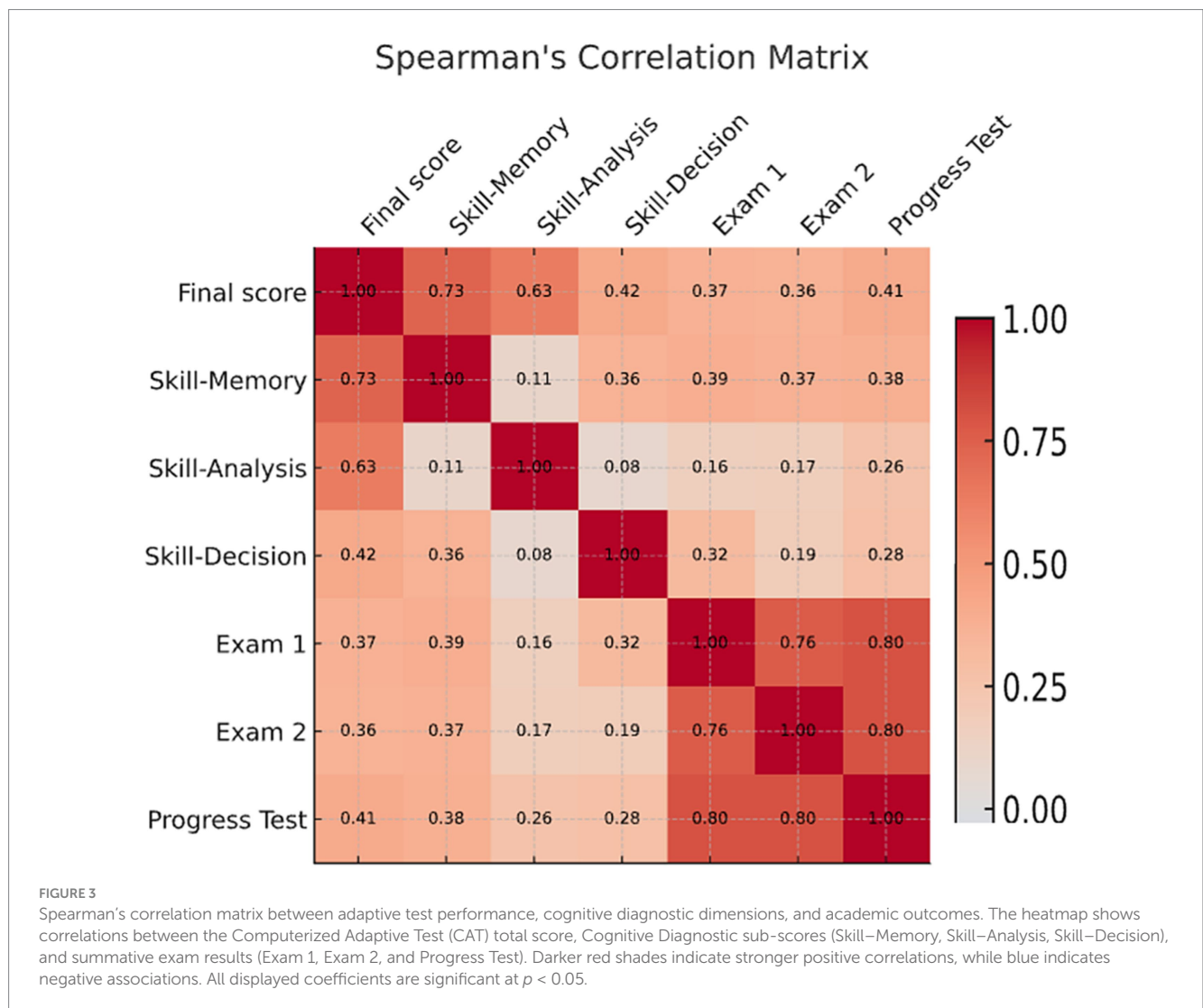
A Mann–Whitney test revealed that CAT participants scored significantly higher on the Progress Test than non-participants ( $U = 1,311$ ,  $p = 0.021$ , rank-biserial correlation =  $-0.271$ ). The differences in Exam 1 ( $U = 1392.5$ ,  $p = 0.054$ ) and Exam 2 did not differ significantly between groups ( $U = 1,466$ ,  $p = 0.115$ ). Figure 4 illustrates the group comparison of Progress Test scores between students who completed the formative CAT and those who did not.

An ANCOVA was conducted to evaluate the impact of participation in the formative CAT exam on the summative score. After adjusting for prior performance (Exam 1 and Exam 2), participation in the formative CAT exam had a statistically significant effect on the Progress Test,  $F(1, 143) = 4.239$ ,  $p = 0.041$ ,  $\eta^2 = 0.029$ . Both Exam 1 ( $F = 35.835$ ,  $p < 0.001$ ) and Exam 2 ( $F = 48.388$ ,  $p < 0.001$ ) were also significant predictors of performance, indicating that earlier academic performance was strongly associated with the outcome. These findings suggest that the formative exam were

TABLE 2 Multiple Linear regression analyses evaluating the predictive value of the final CAT score and cognitive diagnostic sub-scores (Skill–Memory, Skill–Analysis, Skill–Decision) on Progress Test performance.

Predictor	$\beta$ (standardized)	95%CI	$t$	$p$
Skill–Memory	0.41	[0.002, 0.029]	2.22	0.29
Skill–Analysis	0.25	[0.003, 0.021]	1.48	0.143
Skill–Decision	0.05	[–0.008, 0.015]	0.58	0.565
Final CAT score	–0.05	[–0.0025, 0.020]	–0.20	0.845

Model:  $R^2 = 0.23$ , adjusted  $R^2 = 0.20$ ,  $F(4,107) = 7.98$ ,  $p < 0.001$ .



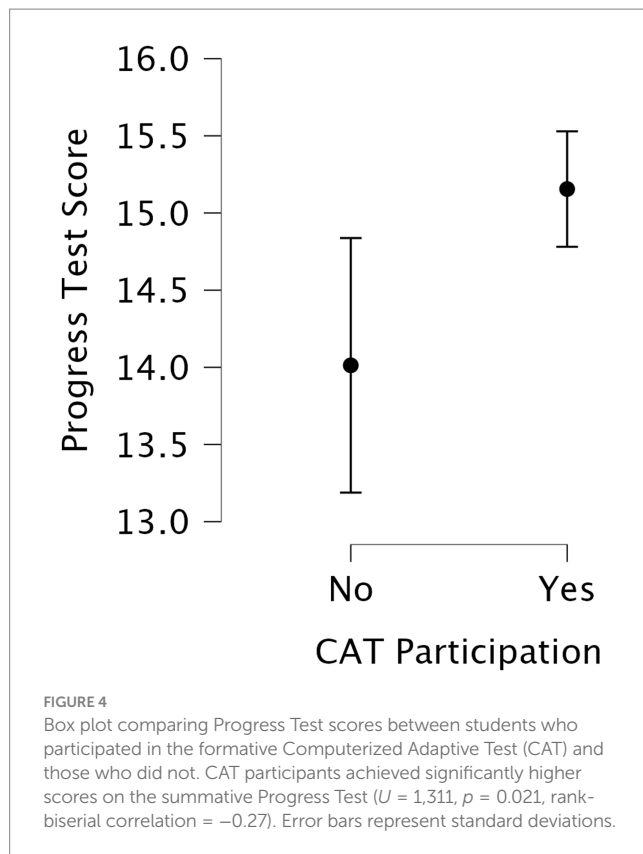
associated with improved performance beyond what could be explained by prior academic achievement alone. Figure 4 illustrates the regression relationships between CAT performance metrics and summative Progress Test results.

A series of linear regressions were conducted to evaluate the predictive value of the final CAT score and each cognitive diagnostic sub-score (Skill–Memory, Skill–Analysis, and Skill–Decision) on Progress Test performance.

The model using the final CAT score as a predictor was statistically significant,  $F(1, 114) = 25.364$ ,  $p < 0.001$ , explaining 18.2% of the variance in Progress Test scores ( $R^2 = 0.182$ ). The final CAT score

emerged as a significant positive predictor ( $\beta = 0.427$ ,  $p < 0.001$ ), indicating that higher adaptive test performance was associated with improved summative outcomes.

When each cognitive skill domain was analyzed separately, Skill–Memory was the strongest predictor of Progress Test performance,  $F(1, 110) = 24.196$ ,  $p < 0.001$ , accounting for 18.0% of the variance ( $R^2 = 0.180$ ;  $\beta = 0.425$ ,  $p < 0.001$ ), indicating a moderate positive association between memory skills and Progress Test performance. Skill–Analysis also significantly predicted Progress Test scores,  $F(1, 110) = 9.574$ ,  $p = 0.003$ , though with a smaller effect size ( $R^2 = 0.080$ ;  $\beta = 0.283$ ,  $p = 0.003$ ). In contrast, Skill–Decision explained only 2.9%



of the variance and was not a significant predictor ( $R^2 = 0.029$ ;  $\beta = 0.170$ ,  $p = 0.072$ ). However, when all domains were entered simultaneously into a multiple regression model, only *Skill-Memory* remained a significant independent predictor ( $\beta = 0.41$ ,  $p = 0.029$ ), confirming its dominant role after controlling for intercorrelations among cognitive domains (Table 1). Figure 5 illustrates the regression relationships between CAT performance metrics and summative outcomes.

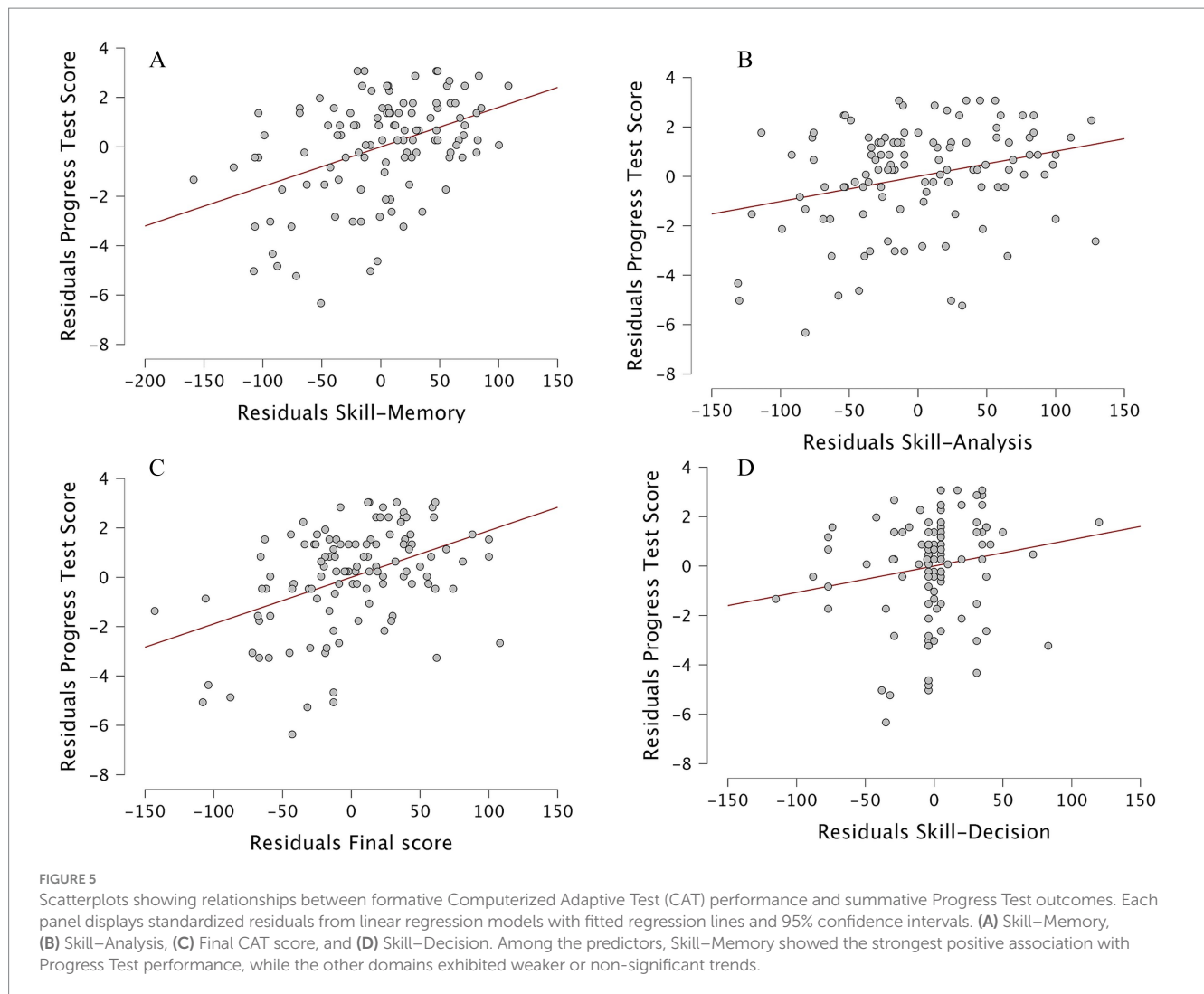
## Discussion

This study explored the use of CAT combined with CDM as a formative assessment tool in undergraduate surgical education. A distinctive feature of the intervention was the integration of AI into both item development and feedback generation, allowing efficient test calibration, individualized difficulty estimation, and automated reporting with minimal faculty workload. These findings align with the growing evidence that AI-driven assessment systems can enhance precision, scalability, and personalization in medical education (Pohn et al., 2025; Shaw et al., 2025; Gordon et al., 2024; Mir et al., 2023; Sunmboye et al., 2025; Gomez et al., 2025).

The association between participation in the formative CAT-CDM exam and higher performance in the subsequent summative Progress Test suggests that adaptive, feedback-oriented assessments may promote more effective study strategies. The multiple regression analysis, including all cognitive domains and the total CAT score, confirmed that only *Skill-Memory* remained a significant independent predictor of summative performance, whereas *Skill-Analysis* and

*Skill-Decision* did not contribute additional explanatory value once intercorrelations were controlled for. This finding indicates that the predictive relationship between analytical and decision-making domains and summative performance is largely shared with memory-based competence. It also reinforces the interpretation that current multiple-choice examination formats primarily reward factual recall rather than complex reasoning or integrative decision-making. These results highlight the need for assessment strategies capable of isolating higher-order cognitive processes from underlying knowledge recall. One possible explanation is that items categorized as decision may not have captured the full complexity of real-world clinical reasoning. As highlighted in previous work, problem-solving in surgery often depends on context, uncertainty, and prioritization rather than discrete knowledge application (Ross et al., 2022; Crebbin et al., 2013). Multiple-choice questions, even when well constructed, are limited in their ability to elicit such integrative reasoning. In contrast, memory items, more closely aligned with the structure of summative exams, may show stronger statistical relationships simply because both assessments rely on similar cognitive processes. This alignment may reflect systemic bias toward factual recall in traditional testing, emphasizing the need for new formats that authentically measure complex reasoning, such as rich clinical vignettes, branching scenarios, or virtual patients (Gomez et al., 2025; Billings et al., n.d.; Rice et al., 2022). Beyond item design limitations, several alternative explanations merit consideration. First, decision-making competence may not yet be sufficiently developed in fifth-year medical students to exhibit measurable variance, given that authentic clinical decision-making typically matures during postgraduate training. Second, the near-zero correlation between analysis and decision domains ( $\rho = 0.08$ ) suggests potential overlap or misalignment within the cognitive classification framework itself, indicating that the current three-domain model may have limited discriminant validity. Third, the summative Progress Test used as the external criterion primarily measures factual and analytical reasoning, which may not adequately fully capture the situational judgment or uncertainty management, considered core features of surgical decision-making. Forth, decision-making items may demand context integration and abstraction beyond what can be effectively assessed in a brief, text-based MCQ, causing students to rely on pattern recognition rather than deliberate reasoning. Fifth, the weak association might also reflect a misalignment between formative and summative constructs: whereas the CAT-CDM aimed to assess applied reasoning, the Progress Test could be predominantly capturing factual recall, thereby reducing shared variance by design. These combined factors likely explain the absence of significant associations and underscore that current multiple-choice formats are inherently constrained in representing complex cognitive processes such as risk-benefit reasoning, ethical trade-offs, and context-specific prioritization. As such, the null finding highlights an important boundary in the construct validity of decision-making assessment and identifies an area for future instrument development and validation.

From an educational perspective, these results have practical implications for curriculum design. Although statistically reliable, the effect sizes were modest, suggesting that the practical impact of short formative interventions may be incremental rather than transformative. If recall remains the main predictor of summative success, surgical educators risk over-emphasizing rote knowledge at the expense of reasoning and decision-making competence.



Integrating adaptive cognitive diagnostics into teaching could help identify students who rely predominantly on memorization and guide them toward deliberate practice in interpretation and judgment. Educators should consider coupling CAT-CDM data with simulation or case-based discussions, transforming feedback into structured remediation plans. In parallel, AI-supported feedback dashboards could provide students with a dynamic map of their evolving cognitive profile, encouraging self-regulated learning and early correction of deficiencies. By integrating these tools, surgical curricula could progressively shift from knowledge reproduction to diagnostic reasoning and decision-making mastery.

Beyond its psychometric contribution, the present study provides several educational insights. The CAT-CDM framework aligns closely with principles of adaptive learning, in which instructional content and assessment dynamically adjust to each learner's ability level, which we believe to be key to learning. By identifying specific cognitive domains requiring reinforcement, adaptive testing provides personalized diagnostic feedback that can guide self-regulated learning. Students can use the domain-specific results to direct their study strategies toward weaker areas, engage in deliberate practice, and monitor progress over time. From an instructional standpoint,

such diagnostic information enables educators to allocate remediation resources more efficiently and to tailor teaching toward common cognitive gaps.

Furthermore, the integration of AI-assisted item generation and feedback represents a scalable model for formative assessment with minimal faculty workload, consistent with the growing emphasis on feedback as a continuous, learner-driven process rather than an episodic event. By converting performance data into interpretable cognitive profiles, this approach helps close the feedback loop and fosters reflection, autonomy, and iterative improvement.

Several limitations must be acknowledged. First, participation in the formative exam was voluntary, introducing potential selection bias, as more motivated students may have been more likely to participate. Although all students from the same curricular unit were invited to participate, and overall participation was high (79%), voluntary participation may have led to self-selection bias, with more motivated or academically stronger students being overrepresented among CAT participants. This possibility should be considered when interpreting the association between formative and summative performance since it restricts the ability to establish causal inferences; therefore, the observed differences should be interpreted as

associations rather than direct effects of the intervention. This short interval of 3–5 days before the summative Progress Test mirrors real-world preparation patterns in which students consolidate learning shortly before assessment. The aim was not to measure long-term knowledge retention but to determine whether adaptive feedback could inform final revision strategies. Future studies will extend the interval between formative and summative assessments to investigate sustained learning effects and behavioral change over time. Second, this was a single-center study, potentially limiting applicability to other educational contexts. Finally, the study focused on short-term outcomes; future research should examine whether the observed benefits persist in long-term knowledge retention and clinical performance. Although initial inter-rater reliability for cognitive classification was substantial ( $\kappa = 0.76$ ), and retrospective double-blind validation of a sample of items demonstrated similarly substantial agreement ( $\kappa = 0.68$ ), some residual subjectivity in distinguishing analysis from decision items cannot be fully excluded. This issue is particularly relevant given the weaker associations observed in the Skill–Decision domain. Future research should include independent expert classification of the entire item bank to further strengthen the construct validity of cognitive domain distinctions.

Future research should employ randomized or crossover designs to confirm efficacy, explore the durability of learning effects, and examine the longitudinal impact of adaptive feedback on self-regulated learning. Establishing causal links between adaptive testing, feedback quality, and performance improvement would provide stronger evidence for scaling this approach. Integrating the CAT–CDM model with simulation-based or virtual-patient environments could enhance authenticity and allow assessment of higher-order decision-making under uncertainty—an essential but often under-evaluated component of surgical competence. Expanding item banks with scenario-driven and branching questions may further strengthen construct validity for complex reasoning. Ultimately, such investigations could contribute to a more holistic and adaptive assessment ecosystem encompassing the cognitive, technical, and non-technical dimensions of surgical education.

## Conclusion

This study found that the use of Computerized Adaptive Testing (CAT) combined with Cognitive Diagnostic Modeling (CDM), supported by AI-based item generation and automated feedback, was associated with higher performance in a subsequent summative assessment. Although decision-making skills were underrepresented in predictive models, the results highlight the need for curricular strategies that better promote higher-order cognitive processing in surgical education. Within the acknowledged methodological limitations, AI-enhanced CAT–CDM emerges as a promising approach for delivering meaningful cognitive feedback and fostering data-informed learning in medical training.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by Ethics Committee for Research in Life and Health Sciences. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

NS: Validation, Resources, Investigation, Visualization, Data curation, Conceptualization, Formal analysis, Project administration, Writing – review & editing, Funding acquisition, Software, Supervision, Methodology, Writing – original draft. CC: Data curation, Validation, Conceptualization, Supervision, Methodology, Project administration, Investigation, Resources, Writing – review & editing, Funding acquisition, Software, Formal analysis, Visualization. JP: Visualization, Resources, Funding acquisition, Validation, Formal analysis, Project administration, Supervision, Data curation, Writing – review & editing, Investigation, Conceptualization, Software, Methodology.

## Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

## Conflict of interest

JP is a co-founder of iCognitus4ALL – IT Solutions (Porto, Portugal), the company that developed QuizOne®, the assessment platform used in this study. This affiliation did not influence study design, data collection, analysis, or interpretation of results. No financial compensation, consultancy fees, or royalties were received by any author for conducting or publishing this research. QuizOne® was accessed through the University of Minho's standard institutional spin-off agreement at no additional cost to the research project, and the platform is routinely employed for both formative and summative assessments across the medical curriculum. iCognitus had no role in manuscript preparation, data analysis, interpretation of findings, or the decision to publish. There were no restrictions on the reporting of results, including null or negative findings.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of *Frontiers*, at the time of submission. This had no impact on the peer review process and the final decision.

## Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by *Frontiers* with the support of artificial

intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the

reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnbeh.2025.1735237/full#supplementary-material>

## References

- Appelhaus, S., Werner, S., Grosse, P., and Kämmer, J. E. (2023). Feedback, fairness, and validity: effects of disclosing and reusing multiple-choice questions in medical schools. *Med. Educ. Online* 28:2143298. doi: 10.1080/10872981.2022.2143298
- Barthakur, A., Kovanovic, V., Joksimovic, S., Zhang, Z., Richey, M., and Pardo, A. (2022). Measuring leadership development in workplace learning using automated assessments: learning analytics and measurement theory approach. *Br. J. Educ. Technol.* 53, 1842–1863. doi: 10.1111/bjet.13218
- Billings, MS, DeRuchie, K, Go, S, Hussie, K, Kulesher, A, and Merrell, J Constructing written test questions for the health SCIENCES n.d.
- Burgess, A., van Diggele, C., Roberts, C., and Mellis, C. (2020). Feedback in the clinical setting. *BMC Med. Educ.* 20:460. doi: 10.1186/s12909-020-02280-5
- Burr, S. A., Gale, T., Kisielewska, J., Millin, P., Pêgo, J. M., Pinter, G., et al. (2016). A narrative review of adaptive testing and its application to medical education. *MedEdPublish* 2023:221. doi: 10.12688/mep.19844.1
- Butler, A. C., and Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Mem. Cogn.* 36, 604–616. doi: 10.3758/MC.36.3.604
- Chang, H.-H., and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Appl. Psychol. Meas.* 20, 213–229. doi: 10.1177/014662169602000303
- Collares, C. F., and Cecilio-Fernandes, D. (2019). When I say ... computerised adaptive testing. *Med. Educ.* 53, 115–116. doi: 10.1111/medu.13648
- Crebbin, W., Beasley, S. W., and Watters, D. A. K. (2013). Clinical decision making: how surgeons do it. *ANZ J. Surg.* 83, 422–428. doi: 10.1111/ans.12180
- Dedy, N. J., Bonrath, E. M., Zevin, B., and Grantcharov, T. P. (2013). Teaching nontechnical skills in surgical residency: a systematic review of current approaches and outcomes. *Surgery* 154, 1000–1008. doi: 10.1016/j.surg.2013.04.034
- El Boghdady, M., and Alijani, A. (2017). Feedback in surgical education. *Surgeon* 15, 98–103. doi: 10.1016/j.surg.2016.06.006
- Falcão, F., Costa, P., and Pêgo, J. M. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Adv. Health Sci. Educ. Theory Pract.* 27, 405–425. doi: 10.1007/s10459-022-10092-z
- Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., and Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of automatic item generation. *Adv. Health Sci. Educ. Theory Pract.* 28, 1441–1465. doi: 10.1007/s10459-023-10225-y
- Garner, M. S., Gusberg, R. J., and Kim, A. W. (2014). The positive effect of immediate feedback on medical student education during the surgical clerkship. *J. Surg. Educ.* 71, 391–397. doi: 10.1016/j.jsurg.2013.10.009
- Gierl, M. J., Shin, J., Firoozi, T., and Lai, H. (2022). Using content coding and automatic item generation to improve test security. *Front. Educ.* 7. doi: 10.3389/educ.2022.853578
- Gomez, C., Seenivasan, L., Zou, X., Yoon, J., Chu, S., Leong, A., et al Explainable AI for automated user-specific feedback in surgical skill acquisition. In *Human-AI collaboration: first international workshop, HAIC 2025, held in conjunction with MICCAI 2025, Daejeon, South Korea 2025*.
- Gordon, M., Daniel, M., Ajiboye, A., Uraiby, H., Xu, N. Y., Bartlett, R., et al. (2024). A scoping review of artificial intelligence in medical education: BEME guide no. 84. *Med. Teach.* 46, 446–470. doi: 10.1080/0142159X.2024.2314198
- Harrison (2017). Feedback in the context of high-stakes assessment: can summative be formative? Maastricht: Maastricht University, doi: 10.26481/dis.20170920ch.
- J. Leighton and M. Gierl (Eds.) (2007). *Cognitive diagnostic assessment for education: theory and applications*. Cambridge: Cambridge University Press doi: 10.1017/CBO9780511611186.
- Lord, F. M. (1971). Robbins-monro procedures for tailored testing. *Educ. Psychol. Meas.* 31, 3–31. doi: 10.1177/001316447103100101
- Ma, C., Ouyang, J., and Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika* 88, 175–207. doi: 10.1007/s11336-022-09867-5
- Madani, A., Vassiliou, M. C., Watanabe, Y., Al-Halabi, B., Al-Rowais, M. S., Deckelbaum, D. L., et al. (2017). What are the principles that guide Behaviors in the operating room? Creating a framework to define and measure performance. *Ann. Surg.* 265, 255–267. doi: 10.1097/SLA.0000000000001962
- Mir, M. M., Mir, G. M., Raina, N. T., Mir, S. M., Mir, S. M., Miskeen, E., et al. (2023). Application of artificial intelligence in medical education: current scenario and future perspectives. *J. Adv. Med. Educ. Prof.* 11, 133–140. doi: 10.30476/JAMP.2023.98655.1803
- Ounounou, E., Aydin, A., Brunckhorst, O., Khan, M. S., Dasgupta, P., and Ahmed, K. (2019). Nontechnical skills in surgery: a systematic review of current training modalities. *J. Surg. Educ.* 76, 14–24. doi: 10.1016/j.jsurg.2018.05.017
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *J. Am. Stat. Assoc.* 70, 351–356. doi: 10.2307/2285821
- Pohn, B., Mehnen, L., Fitzek, S., Choi, K.-E., Braun, R. J., and Hatamikia, S. (2025). Integrating artificial intelligence into pre-clinical medical education: challenges, opportunities, and recommendations. *Front. Educ.* 10. doi: 10.3389/educ.2025.1570389
- Rice, N., Pêgo, J. M., Collares, C. F., Kisielewska, J., and Gale, T. (2022). The development and implementation of a computer adaptive progress test across European countries. *Comput. Educ.* 3:100083. doi: 10.1016/j.caeai.2022.100083
- Rosendal, A. A., Sloth, S. B., Rölfing, J. D., Bie, M., and Jensen, R. D. (2023). Technical, non-technical, or both? A scoping review of skills in simulation-based surgical training. *J. Surg. Educ.* 80, 731–749. doi: 10.1016/j.jsurg.2023.02.011
- Ross, S., Pirraglia, C., Aquilina, A. M., and Zulla, R. (2022). Effective competency-based medical education requires learning environments that promote a mastery goal orientation: a narrative review. *Med. Teach.* 44, 527–534. doi: 10.1080/0142159X.2021.2004307
- Seo, D. G., Choi, J., and Kim, J. (2024). Comparison of real data and simulated data analysis of a stopping rule based on the standard error of measurement in computerized adaptive testing for medical examinations in Korea: a psychometric study. *J. Educ. Eval. Health Prof.* 21:18. doi: 10.3352/jeehp.2024.21.18
- Shaughness, G., Georgoff, P. E., Sandhu, G., Leininger, L., Nikolian, V. C., Reddy, R., et al. (2017). Assessment of clinical feedback given to medical students via an electronic feedback system. *J. Surg. Res.* 218, 174–179. doi: 10.1016/j.jss.2017.05.055
- Shaw, K., Henning, M. A., and Webster, C. S. (2025). Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. *Med. Sci. Educ.* 35, 1803–1816. doi: 10.1007/s40670-025-02373-0
- Sunmboye, K., Strafford, H., Noorestani, S., and Wilson-Pirie, M. (2025). Exploring the influence of artificial intelligence integration on personalized learning: a cross-sectional study of undergraduate medical students in the United Kingdom. *BMC Med. Educ.* 25:570. doi: 10.1186/s12909-025-07084-z
- Van Wijk, E. V., Donkers, J., De Laat, P. C. J., Meiboom, A. A., Jacobs, B., Ravesloot, J. H., et al. (n.d.). Computer adaptive vs. non-adaptive medical Progress testing: feasibility, test performance, and student experiences. *Perspect Med Educ* 13, 406–416. doi: 10.5334/pme.1345
- Williamson, J. Cognitive diagnostic models and how they can be useful n.d.
- Xu, L., Jiang, Z., Han, Y., Liang, H., and Ouyang, J. (2023). Developing computerized adaptive testing for a national health professionals exam: an attempt from psychometric simulations. *Perspect. Med. Educ.* 12, 462–471. doi: 10.5334/pme.855