Original Paper

Learning Clinical Reasoning With Virtual Patients Using Adaptive vs Static Feedback in Individual and Collaborative Learning Modes: Randomized Controlled Trial

Jan Kiesewetter[1], Prof Dr; Helene Werl[1], MD; Katharina Nederstigt[1], MD; Sven Sarbu-Rothsching[1], MD; Michael Sailer[2], Prof Dr; Elisabeth Bauer[2], PhD; Jan M Zottmann[1], PhD; Frank Fischer[3], Prof Dr; Martin R Fischer[1], Prof Dr

[1]Institute of Medical Education, LMU Medizin, LMU University Hospital, München, Germany
[2]Learning Analytics and Educational Data Mining, University of Augsburg, Augsburg, Germany
[3]LMU München, Lehrstuhl für Pädagogische Psychologie und Empirische Pädagogik, München, Germany

Corresponding Author:

Jan Kiesewetter, Prof Dr
Institute of Medical Education, LMU Medizin, LMU University Hospital
Pettenkoferstr. 8a
München 80336
Germany
Phone: +49 89 4400 ext 57213
Email: jan.kiesewetter@med.lmu.de

Abstract

Background: Learning clinical reasoning with virtual patients can be supported by feedback and collaboration, but little is known about how the type of feedback and the learning mode interact.

Objective: This study examined the effects of adaptive vs static feedback and of individual vs collaborative learning on medical students' conceptual knowledge, strategic knowledge, and diagnostic accuracy when learning clinical reasoning with virtual patients.

Methods: In a pre-post 2 × 2 factorial laboratory study, 103 medical students (third- to sixth-year) worked on 8 virtual patients either individually (n=31) or collaboratively in dyads (n=72; 36 dyads). Individuals or dyads were randomly assigned to receive either static feedback (a written expert solution) or automated adaptive feedback based on a natural language processing analysis of their written diagnostic explanations. Conceptual knowledge (multiple-choice items) and strategic knowledge (key feature problems) were assessed directly before and 1 week after the learning phase. Diagnostic accuracy was assessed during the learning phase (immediate effect) and with 2 new virtual patients in the delayed posttest (durability of the effect).

Results: Students in all groups improved their conceptual knowledge over time ($F_{1,97}$=393.17; P<.001; η^2=0.80). Students who received adaptive feedback showed higher conceptual knowledge ($F_{1,97}$=4.87; P=.03; η^2=0.048) and higher strategic knowledge ($F_{1,109}$=4.40; P=.04) in the delayed posttest than students who received static feedback, whereas the type of feedback did not affect diagnostic accuracy ($F_{1,97}$=0.23; P=.63). Students who learned collaboratively displayed higher diagnostic accuracy than individual learners during the learning phase ($F_{1,83}$=10.17; P=.002; η^2=0.11), but not in the delayed posttest, and they retained strategic knowledge better than individual learners ($F_{1,133}$=7.27; P=.008; η^2=0.052). No significant interaction between the type of feedback and learning mode was found ($F_{1,131}$=0.07; P=.80).

Conclusions: Adaptive feedback promoted the retention of conceptual and strategic knowledge in both learning modes but did not affect diagnostic accuracy. The advantage of collaborative learning for diagnostic accuracy was limited to the phase in which students actually worked together and did not persist when students were tested individually 1 week later. Adaptive feedback and collaborative learning thus support different components of learning clinical reasoning and can be combined in virtual patient environments.

JMIR Med Educ 2026;12:e77606; doi: 10.2196/77606

Keywords: clinical reasoning; virtual patients; adaptive feedback; collaborative learning; diagnostic accuracy

Introduction

Background

Clinical reasoning is a key skill of the medical profession. Therefore, it is necessary to teach and practice clinical reasoning during medical education. In recent years, students have increasingly used so-called virtual patients to develop such skills [1]. There are numerous variations and forms of virtual patients, and insights have been gained into how virtual patients need to be designed for medical students to learn effectively with them [2]. We define and operationalize virtual patients as simple digital case presentations, where the learner prioritizes tasks in order to make diagnostic and/or therapeutic decisions [1,2]. A recent study has shown that students need to have a certain amount of prior knowledge in order to benefit from working with virtual patients [3]. Other findings suggest that it is possible to learn aspects of clinical reasoning with virtual patients even with little prior knowledge, when proper guidance is provided in the form of reflective practice, a specific scaffolding practice, or feedback [4].

In many virtual patient environments, students enter diagnoses, and all students receive the same feedback with an explanation of why a certain diagnosis is considered correct. While meta-analyses highlight the benefits of feedback on the current state of performance and information on how to improve future task processing (feed-forward) [5,6], many virtual patient systems provide only limited feedback. Process-related feedback needs to move beyond the accuracy of a concluded diagnosis and requires data on the reasoning processes in the virtual patient environment. For example, self-explanation prompts or justifications require students to write responses explaining why they believe a certain diagnosis might be correct [7]. Based on these answers, instructors can provide adaptive feedback tailored to the prior knowledge and current performance of the respective student. Instructors can provide feedback on the parts of a student's answer or diagnosis that are correct and provide additional guidance for the parts that are incorrect or incomplete. Since providing such feedback individually to students working with virtual patients is highly resource-demanding, we propose computer-supported adaptive feedback as an interactive, resource-optimized, and scalable alternative [8]. We previously implemented adaptive feedback in the context of teacher education and found that it helped improve the quality of justifications, but not diagnostic accuracy in individual learners [8]. Diagnostic accuracy—whether the diagnosis entered by students is correct—is one of the key outcomes of any clinical reasoning endeavor [9]. Static feedback even had detrimental effects on learning processes when students were diagnosing a classroom situation collaboratively in dyads [8]. However, in medical education, not a classroom situation but rather an individual virtual patient is diagnosed, and objective medical tests and examinations can be performed, providing a source of feedback within the reasoning process. Taking this difference between the medical and teacher domains into consideration, we aimed to compare different types of feedback (ie, automated adaptive feedback and static feedback) in the context of medical education.

There are 2 main learning modes: collaborative and individual learning [9]. Some studies have shown that learners working in dyads may achieve higher diagnostic accuracy than individual learners, possibly because collaboration allows learners to generate and evaluate diagnostic hypotheses jointly [9]. However, working collaboratively is typically accompanied by a form of feedback from the learning partner and adds complexity to handling the virtual patient. This complexity might hinder diagnostic performance and potentially impede learning. Adaptive feedback tailored to learning partners could help guide the diagnostic process such that high diagnostic performance and learning are achieved. Our study in the context of teacher education points in the direction that collaborative learners might benefit more from adaptive feedback, compared with static feedback, than individual learners [8]. However, little is known about whether these findings can be generalized to medical students' learning.

We thus aim to compare how different types of feedback (adaptive vs static feedback) and learning modes (individual vs collaborative) interact as medical students learn clinical reasoning with virtual patients.

Theoretical Framework

Learning clinical reasoning with virtual patients involves complex cognitive processes that require learners to generate, evaluate, and revise diagnostic hypotheses based on clinical information. The present study builds on theoretical perspectives from feedback research and collaborative learning to explain how instructional support may influence learning outcomes in such environments.

Feedback research suggests that information about the current state of performance and guidance on how to improve future performance can significantly enhance learning outcomes [5,6]. According to the feedback model proposed by Hattie and Timperley [6], effective feedback should address not only task performance but also the processes underlying task completion. In diagnostic reasoning tasks, process-oriented feedback may therefore be particularly important because learners must coordinate multiple cognitive activities such as hypothesis generation, evidence evaluation, and drawing conclusions. Adaptive feedback that responds to learners' explanations may therefore support the development of diagnostic reasoning more effectively than static feedback that provides identical information to all learners.

In addition to feedback, the mode of learning may influence how learners process clinical cases. Research on collaborative learning suggests that working in dyads can enhance performance on complex tasks by allowing learners to pool cognitive resources and jointly construct explanations [10]. At the same time, collaboration may increase task

complexity and cognitive load, as learners must coordinate their reasoning processes and communicate their ideas. The cognitive load perspective on collaborative learning, therefore, suggests that collaboration can either facilitate or hinder learning depending on the instructional support provided.

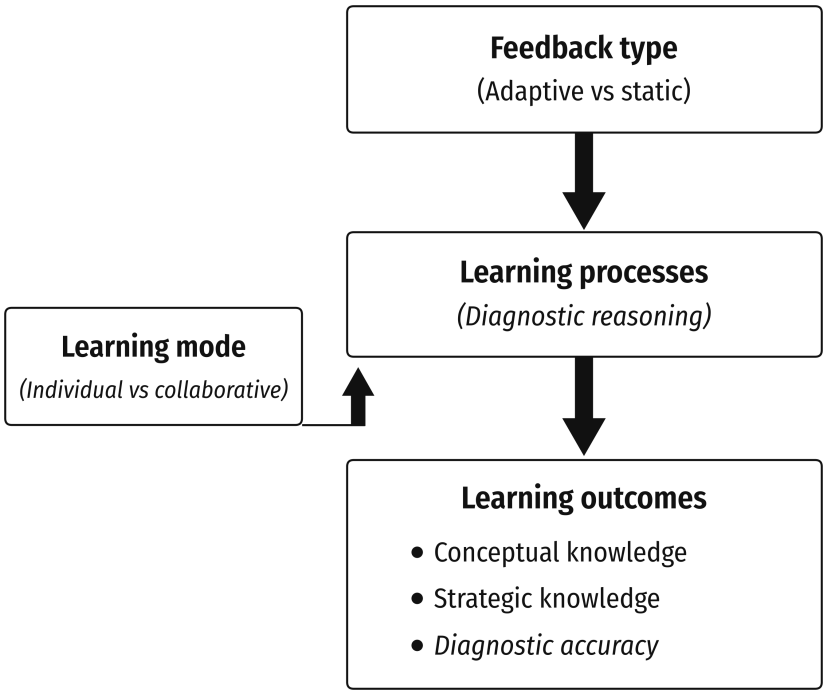
Diagnostic reasoning tasks in medical education represent a particularly complex form of problem-solving that requires integrating prior knowledge with case-specific information. Previous research has shown that collaborative diagnosis can lead to higher diagnostic accuracy compared with individual reasoning, particularly when learners generate and discuss alternative hypotheses [11]. In such situations, adaptive feedback may be especially beneficial because it can guide the reasoning processes of individuals or dyads and help them regulate their diagnostic activities.

From a social constructivist perspective [12], collaborative learning can facilitate knowledge construction through explanation, argumentation, and negotiation of meaning. When learners jointly discuss diagnostic hypotheses, they may benefit from distributed cognitive resources and mutual scaffolding. At the same time, collaboration can increase cognitive load because learners must coordinate their reasoning processes and communication. Therefore, cognitive load theory suggests that instructional support such as adaptive feedback may be particularly beneficial in collaborative environments by guiding attention toward diagnostically relevant information.

Beyond diagnostic accuracy, acquiring conceptual and strategic knowledge is essential for developing clinical reasoning. In this study, conceptual knowledge refers to the understanding of medical facts, concepts, and their relationships that are relevant to diagnosing diseases. Strategic knowledge involves the application of diagnostic strategies, such as selecting appropriate tests or evaluating alternative diagnoses. Strategic knowledge can, for example, be assessed using key feature problems that capture critical decision-making points in the clinical reasoning process.

Based on these theoretical considerations, the present study examines how 2 instructional design features—type of feedback (adaptive vs static) and learning mode (individual vs collaborative)—interact when medical students learn clinical reasoning with virtual patients (Figure 1).

Figure 1. Theoretical framework of the current study.



The present study addresses the following research questions:

Research question 1: To what extent do medical students benefit from adaptive (compared with static) feedback when learning clinical reasoning with virtual patients?

Research question 2: To what extent do the effects of adaptive (compared with static) feedback depend on the learning mode (individual vs collaborative) when learning clinical reasoning with virtual patients?

Based on prior research on feedback and collaborative learning, we expected that adaptive feedback would lead to higher conceptual and strategic knowledge compared with static feedback. Furthermore, based on findings suggesting the benefits of collaborative diagnostic reasoning, we expected that students working collaboratively would show higher diagnostic accuracy than individual learners. Finally, we explored whether the effectiveness of adaptive feedback would depend on the learning mode. Understanding these relationships can inform the design of scalable instructional support in virtual patient learning environments.

Methods

Sample and Study Design

Participants were third- to sixth-year medical students of LMU Munich (total sample of N=105, mean 24, SD 3.3 y, 69 (66%) female participants; 2 students ultimately had to be excluded from the analyses, resulting in a final sample size of N=103). At the time of recruitment, approximately 1500 students were enrolled in the third to sixth year of the medical program at LMU Munich and were therefore eligible to participate. All student responses used for natural language processing (NLP) analysis were pseudonymized before processing. The system analyzed only textual explanations provided within the learning platform, and no personal identifiers were included in the analysis. Data processing complied with institutional data protection regulations.

Our study used a pre-post, 2 × 2 factorial design with the interindividual factors of type of feedback (static vs adaptive) and learning mode (individual vs collaborative), and the intraindividual factor of time of measurement (prelearning phase, learning phase, and a postlearning phase 1 week later).

Students' knowledge was assessed directly before and 1 week after working on a total of 8 virtual patients (learning phase), either individually (n=33) or collaboratively (n=72 students and n=36 dyads).

Ethical Considerations

Participation in the study was voluntary, and recruitment was conducted via email invitations sent to this population. Ethics approval was granted by the Ethics Committee of the Medical Faculty of LMU Munich (study 17-250). Participants received compensation of €35 (approximately US $40 as of November 2026) for their time. All virtual patients provided were new to all participants. All participants gave informed consent.

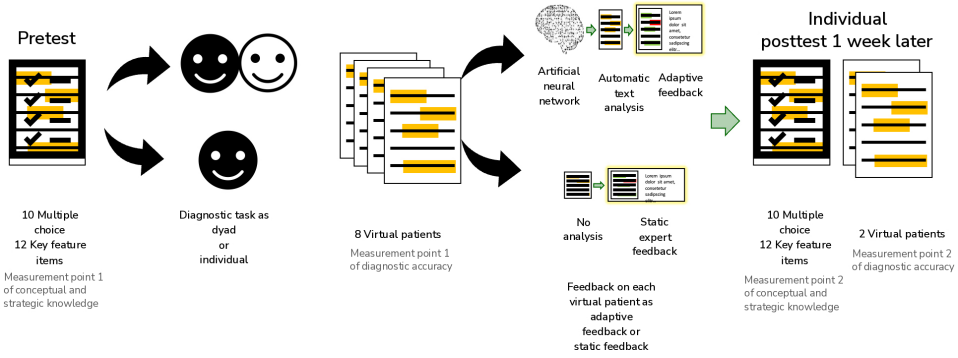
Independent Variables

Based on the date of the test, study participants were assigned to the individual learning mode or to a dyad (ie, the collaborative learning mode). Subsequently, every individual or dyad was randomly assigned to either static or adaptive feedback conditions. Thus, allocation to the type of feedback was fully automated and random, whereas allocation to the learning mode was constrained by the laboratory scheduling and the availability of dyad partners and was therefore not random at the individual level. This allocation procedure also explains the unequal group sizes between the individual and collaborative learning modes.

Procedure

Participants took part in the study in a computer lab at LMU Munich. The procedure was split into a pretest, followed by a learning phase, and a posttest was eventually conducted 1 week later (Figure 2).

Figure 2. Procedure of the study. Conceptual and strategic knowledge were measured at 2 points: in the pretest and in the delayed posttest conducted 1 week later. Diagnostic accuracy was measured twice: during the learning phase (immediate effect; 8 virtual patients with feedback) and in the delayed posttest (durability of the effect; 2 new virtual patients without feedback).



After a brief introduction, participants provided informed consent and logged in to the learning system CASUS (Instruct gGmbH) [13]. The CASUS platform presents virtual patient cases primarily in a text-based interactive format. Students are provided with an initial patient vignette and can subsequently request additional clinical information, such as medical history, physical examination findings, laboratory results, electrocardiogram (ECG) data, or radiological imaging. The interaction, therefore, resembles a structured clinical case exploration rather than a simulation with avatars or video interaction. Students select which diagnostic information to access and in what sequence before entering a final diagnosis and written explanation.

After logging in, students worked on the pretest individually. The pretest was a knowledge test consisting of 10 multiple-choice items and 4 key feature problems comprising 12 items.

During the learning phase, students worked on 8 virtual patients (all materials can be found on the Open Science Framework (OSF) [14], and the cases were initially validated by experts [3] with the instruction to diagnose the patient and

provide a reasoning for their decision. The 8 virtual patient cases represented common diagnostic scenarios in internal medicine. Four cases presented patients with the leading symptom of back pain, and 4 cases presented patients with fever. Each case had a distinct underlying diagnosis (eg, infectious, inflammatory, or mechanical causes), allowing students to practice differential diagnostic reasoning. The cases were designed to reflect realistic clinical presentations while maintaining comparable complexity across cases. All cases were previously validated by medical experts and used in earlier studies investigating learning with virtual patients [3].

According to the experimental condition to which they were assigned, the students worked either individually or collaboratively in dyads. In the CASUS system, students could access different types of clinical information related to the respective virtual patients (eg, laboratory results, ECG, and radiological imaging results). Students could decide how much clinical information they wanted to view and in which order. After submitting their final diagnosis for a virtual patient, students were asked to write down an explanation for their suggested diagnosis. After that, students received either automated adaptive or static feedback on their reasoning (see Bauer et al [15] for a detailed description of the learning environment and Kiesewetter et al [16] for the collaborative learning environment). After 4 of the 8 virtual patients, students had a 5-minute break.

One week after the intervention, all participants worked individually on the posttest. The posttest was a knowledge test that also consisted of 10 multiple-choice items and 12 key feature items and additionally included 2 more virtual patients, on which all participants worked individually. The implementation of such a time-delayed posttest allows the extent to which an intervention leads to sustainable effects to be examined.

Types of Feedback

Static Feedback

Students received a written expert solution after each virtual patient and were asked to compare it with their own solution. Four domain experts from internal medicine validated the expert solutions independently prior to their use in this study. In cases of incoherent expert solutions, a discussion was stipulated until agreement on the expert solution was found. This was only necessary in one of the cases.

Adaptive Feedback

The students' explanations of their diagnoses were analyzed in 2 ways. First, the keyword-based analysis of medical content was implemented using a rule-based algorithm that searched students' responses for predefined medical keywords and keyword combinations associated with the correct diagnostic reasoning steps. This approach allowed the system to identify whether key diagnostic elements were present in the explanation. Tolerance was given to spelling errors so that keywords could be found even if they were spelled incorrectly. The medical content feedback was formulated in a way that allowed for feedback and explanation of correct and incorrect answers, and was written by the medical experts, analogous to the static feedback in length and form, and presented to the students if the relevant keywords could not be found or wrong keywords were found.

Second, the students' reasoning was analyzed using NLP. In order to provide students with adaptive elaborated feedback on their diagnostic reasoning, a learning environment consisting of 3 components (NeuralWeb, INCEpTION, and CASUS) was implemented [17] (Figure 3). Before applying the feedback in the current study, we needed to train the NLP model. We asked 5 domain experts to sequence and code data from 142 individual medical students for so-called epistemic activities (eg, hypothesis generation, evidence generation, evidence evaluation, and drawing conclusions). Interrater reliability was calculated using macro F_1-scores, indicating substantial agreement (minimal upper bound=85.58, maximal upper bound=90.25) [17]. The experts used INCEpTION [18] to annotate the data for the epistemic activities (see [19] for details). Then, based on the coded data, the predictive NLP model was trained in the Python-based web service NeuralWeb. NeuralWeb is a supervised machine learning classifier; it does not use a large language model or retrieval-augmented generation. Its predictions are based exclusively on the expert-annotated training data described above: incoming student explanations were segmented and classified into epistemic activities, and each classification triggered a corresponding preformulated feedback paragraph. The students' answers were processed through the NeuralWeb model to identify which epistemic activities students showed in comparison to the expert model. The resulting feedback was then automatically presented to the students (Figure 4).

Figure 3. Overview of the generation of automated adaptive feedback in our system. In a previous sample, epistemic activities of students working on virtual patients were annotated using INCEpTION, an open-source tool for the annotation of text data (lower left corner). These data were used in NeuralWeb, a Python-based web service for training and running the natural language processing (NLP) model, to train a predictive model and generate a label set of discrete diagnostic classes for the epistemic activities (top center). These labels were then used to provide feedback on students' answers in CASUS, the virtual patient platform in which the students worked on the cases (lower right corner).

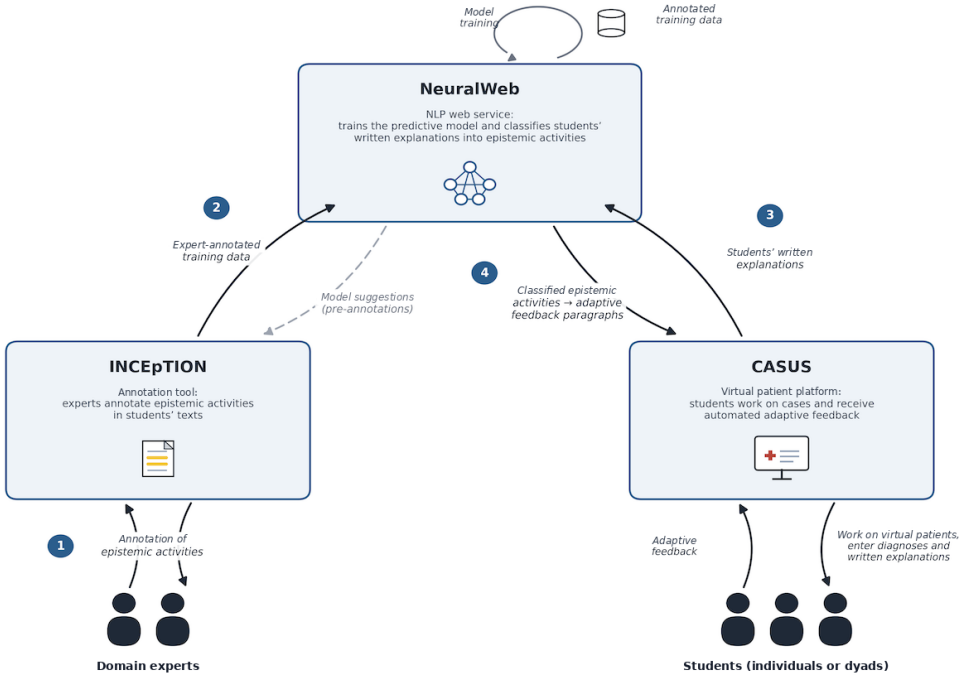
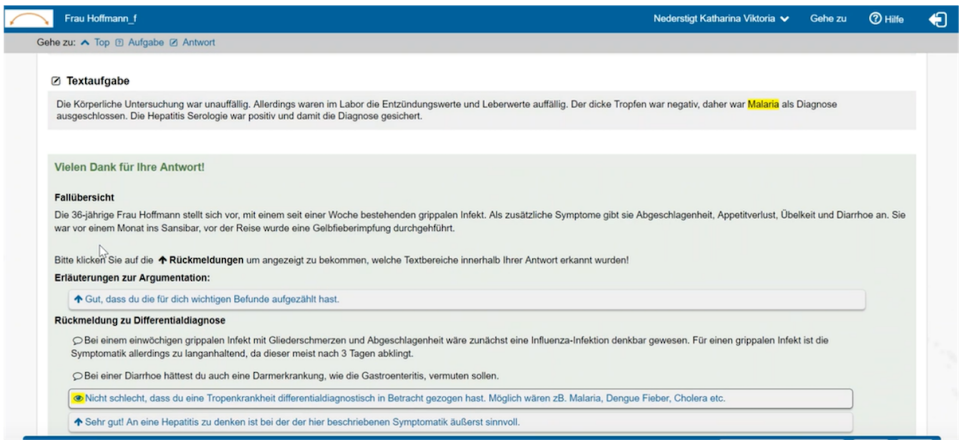


Figure 4. Students' explanation with adaptive feedback. The highlighted text shows which part of the students' explanation triggered the feedback.

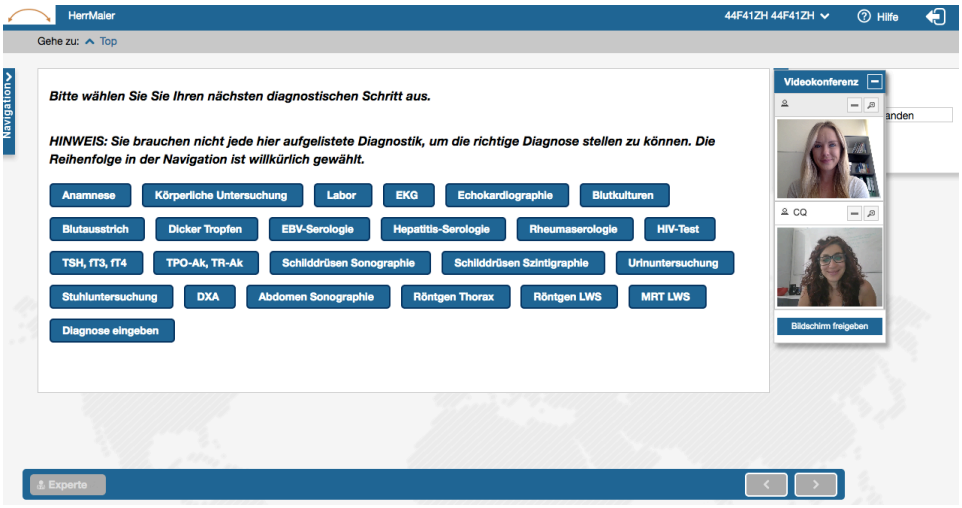


For example, these adaptive elaborated feedback paragraphs informed the students whether they had provided alternative hypotheses or evaluated appropriate evidence in their written reasoning. The paragraphs were provided as feedback on the CASUS platform, together with feedback on the medical content. Students were able to view which parts of their reasoning triggered which parts of the automated feedback by clicking on the feedback paragraphs, which then highlighted the respective text in their original answers.

Learning Modes

The students working individually were assigned a place in the computer laboratory. Students in the collaborative learning mode were placed in separate computer laboratory rooms and additionally provided with a headset (Figure 5). For technical details on the collaborative learning environment, see the study by Bauer et al [15]. The learning environment was the same, but the students entered a code at the beginning of the test, which enabled a connection to their learning partner via video chat. One student shared their screen and controlled the keyboard and mouse, while the other student watched the shared screen to discuss the virtual patient and next steps with their learning partner. After half of the virtual patients were completed, control of the mouse and keyboard switched to the other student in the dyad. Although both students in a dyad participated at the same time, they were seated in separate rooms and collaborated exclusively via this remote collaboration interface (video chat and screen sharing). This setup was chosen to standardize communication across all dyads, to ensure that the entire collaboration was mediated by the learning environment, and to keep the interface identical to the previously evaluated remote collaboration platform [16].

Figure 5. Screenshot of the learning environment in collaborative learning mode.



Dependent Variables

Assessment of Knowledge

We developed a knowledge test that consisted of 10 multiple-choice items to assess conceptual knowledge. Additionally, strategic knowledge was assessed with 4 key feature problems that comprised 12 items (3 items per problem). The test items covered the medical content relevant to working with the virtual patients (ie, diseases presenting with back pain or fever). To avoid a retest effect, the knowledge test was piloted on a sample of the study population and divided into 2 parallel tests, creating paired items for the knowledge pretest and the knowledge posttest based on difficulty and reliability.

The knowledge test was developed specifically for the present study by a team of medical educators and domain experts in internal medicine. The test focused on clinical knowledge relevant to the virtual patient cases, particularly diagnostic reasoning related to common causes of back pain and fever. The items were designed to assess 2 domains of knowledge: conceptual medical knowledge and strategic knowledge relevant to clinical decision-making. All items were reviewed by 4 medical experts for content validity before use.

The knowledge test was conducted before and 1 week after the learning phase. Scores were reported and analyzed as the percentage of correct responses. The reliability in the validation of the knowledge test was Cronbach α=0.64 for the knowledge pretest overall and α=0.65 for the knowledge posttest. In multivariate testing, no differences in the interindividual variables, feedback, and learning mode were found in the pretest for conceptual and strategic knowledge.

Diagnostic Accuracy

Students entered their diagnoses as open-text responses in the learning system. These responses were subsequently coded by comparing them to the predefined correct diagnoses. Only the correct final diagnosis was scored as correct. Responses that referred to related but nonspecific categories (eg, “infection” instead of a specific diagnosis) were coded as incorrect to ensure consistent scoring criteria. For each correctly identified diagnosis, students received 1 point. A false diagnosis resulted in zero points. The number of correct diagnoses was used as an indicator of diagnostic accuracy (range 0-8 points). For the assessment of diagnostic accuracy in the posttest, 2 additional new virtual patients were analyzed individually by the participants. None of the participants received feedback when working on the virtual patients during the posttest.

Data Analysis

Data were analyzed using SPSS 26 (IBM Corp). All data were analyzed in repeated-measures ANOVAs with time of measurement (pretest and posttest), type of feedback (static vs adaptive), and learning mode (individual vs collaborative) as independent variables. To stratify for time-on-task, we incorporated time-on-task as a covariate in all repeated-measures ANOVAs.

We used listwise elimination of missing data. Therefore, 2 students who worked individually with the virtual patients had to be excluded from the sample due to technical difficulties. Interpretation of partial η^2 is according to Cohen [20] (throughout, η^2 denotes partial η^2). Post hoc analyses were least significant difference tests. The α level was set at P<.05.

Results

Type of Feedback: Static vs Adaptive Feedback

With respect to conceptual knowledge, we found a large main effect of time of measurement ($F_{1,97}$=393.17; P<.001; η^2=0.80). We found a small interaction effect of time of measurement and feedback ($F_{1,97}$=4.865; P=.03; η^2=0.048). Students who received adaptive feedback outperformed students who received static feedback (Table 1).

Table 1. Overview of the results for conceptual knowledge.

Time of measurement, type of feedback, and learning mode	Proportion correct (SD)
Pretest (multiple-choice)	
Static	
Individual	0.46 (0.15)
Collaborative	0.43 (0.11)
Overall	0.44 (0.13)
Adaptive	
Individual	0.42 (0.18)
Collaborative	0.39 (0.12)
Overall	0.41 (0.15)
Overall	
Individual	0.44 (0.16)
Collaborative	0.42 (0.11)
Overall	0.43 (0.14)
Posttest (multiple-choice)	
Static	
Individual	0.74 (0.11)
Collaborative	0.68 (0.12)
Overall	0.70 (0.12)
Adaptive	
Individual	0.78 (0.17)
Collaborative	0.78 (0.11)
Overall	0.78 (0.14)
Overall	
Individual	0.76 (0.14)
Collaborative	0.72 (0.12)
Overall	0.74 (0.13)

With respect to strategic knowledge, we found a large main effect for time of measurement ($F_{1,109}$=19.663; P<.001; η^2=0.15) and a small-sized interaction effect between time of measurement and feedback ($F_{1,109}$=4.397; P=.04; η^2=0.039). Students with adaptive feedback outperformed those with static feedback in the posttest, but not in the pretest (Table 2).

Table 2. Overview of the results for strategic knowledge.

Time of measurement, type of feedback, and learning mode	Proportion correct (SD)
Pretest (key feature)	
Static	
Individual	0.64 (0.09)
Collaborative	0.67 (0.05)
Overall	0.66 (0.07)
Adaptive	
Individual	0.67 (0.07)
Collaborative	0.67 (0.04)
Overall	0.67 (0.06)
Overall	
Individual	0.65 (0.08)
Collaborative	0.67 (0.05)
Overall	0.66 (0.07)
Posttest (key feature)	
Static	
Individual	0.48 (0.28)

Time of measurement, type of feedback, and learning mode	Proportion correct (SD)
Collaborative	0.59 (0.15)
Overall	0.54 (0.23)
Adaptive	
Individual	0.61 (0.10)
Collaborative	0.64 (0.08)
Overall	0.62 (0.09)
Overall	
Individual	0.54 (0.23)
Collaborative	0.61 (0.13)
Overall	0.57 (0.19)

Diagnostic Accuracy

We found a small main effect of time of measurement on diagnostic accuracy ($F_{1,97}$=5.18; P=.03; η^2=0.05). Independent of the type of feedback and learning mode, the students had higher diagnostic accuracy on the posttest 1 week after completing 8 virtual patients. We found no main effect of type of feedback on diagnostic accuracy ($F_{1,97}$=0.232; P=.63; not significant).

Learning Mode: Individual vs Collaborative Learning

With respect to conceptual knowledge, no effects were found for the time of measurement and learning mode ($F_{1,97}$=0.21; P=.65; not significant; see Table 1).

With respect to strategic knowledge, we found a small interaction effect between time of measurement and learning mode ($F_{1,133}$=7.27; P=.008; η^2=0.052). Independent of the type of feedback given, students who learned collaboratively in dyads outperformed individual learners (Table 2).

We found a medium-sized main effect of the learning mode on diagnostic accuracy in the learning phase ($F_{1,83}$=10.17; P=.002; η^2=0.11). Learners who learned in the collaborative learning mode displayed higher diagnostic accuracy (mean 0.56, SD 0.17) than individual learners (mean 0.46, SD 0.20). No analogous effect was found for diagnostic accuracy in the posttest (Table 3). We found no interaction between the type of feedback and the learning mode ($F_{1,131}$=0.07; P=.80; not significant).

Table 3. Overview of the results for diagnostic accuracy in percent.

Time of measurement, type of feedback, and learning mode	Proportion correct (SD)
During learning phase	
Static	
Individual	0.49 (0.18)
Collaborative	0.58 (0.17)
Total	0.56 (0.18)
Adaptive	
Individual	0.43 (0.21)
Collaborative	0.50 (0.16)
Total	0.47 (0.18)
Total	
Individual	0.46 (0.20)
Collaborative	0.56 (0.17)
Total	0.52 (0.19)
Posttest	
Static	
Individual	0.62 (0.22)
Collaborative	0.62 (0.33)
Total	0.62 (0.30)
Adaptive	
Individual	0.66 (0.34)
Collaborative	0.66 (0.24)
Total	0.66 (0.28)
Total	

Time of measurement, type of feedback, and learning mode	Proportion correct (SD)
Individual	0.64 (0.28)
Collaborative	0.64 (0.30)
Total	0.64 (0.29)

Discussion

Principal Findings

In this 2 × 2 factorial pre-post study, students in all groups improved their conceptual knowledge from pretest to posttest. Beyond this general learning effect, the type of feedback had small but significant effects on knowledge retention: students who received adaptive feedback showed higher conceptual and strategic knowledge in the delayed posttest than students who received static feedback, whereas the type of feedback did not affect diagnostic accuracy. The learning mode affected performance differently: students working collaboratively in dyads achieved higher diagnostic accuracy than individual learners while they were actually working together during the learning phase, but this advantage did not persist in the delayed posttest, in which all students worked individually. In addition, collaborative learners retained strategic knowledge better than individual learners. No significant interaction between type of feedback and learning mode was found. Table 4 links the hypotheses to the corresponding outcome measures and results.

Table 4. Overview of hypotheses, outcome measures, and corresponding results.

Hypothesis	Outcome measure	Result	Conclusion
H1: Adaptive feedback leads to higher knowledge than static feedback	Conceptual knowledge (multiple-choice items)	Time × feedback interaction; $F_{1,97}$=4.87; P=.03; η^2=0.048; adaptive>static in the posttest	Supported
H1: Adaptive feedback leads to higher knowledge than static feedback	Strategic knowledge (key feature problems)	Time × feedback interaction; $F_{1,109}$=4.40; P=.04; adaptive>static in the posttest	Supported
H1: Adaptive feedback leads to higher knowledge than static feedback	Diagnostic accuracy	No main effect of type of feedback; $F_{1,97}$=0.23; P=.63	Not supported
H2: Collaborative learners show higher diagnostic accuracy than individual learners	Diagnostic accuracy during the learning phase	Main effect of learning mode; $F_{1,83}$=10.17; P=.002; η^2=0.11; collaborative>individual	Supported
H2: Collaborative learners show higher diagnostic accuracy than individual learners	Diagnostic accuracy in the delayed posttest (individual testing)	No effect of learning mode	Not supported
H2: Collaborative learners show higher diagnostic accuracy than individual learners	Strategic knowledge (key feature problems)	Time × learning mode interaction $F_{1,133}$=7.27; P=.008; η^2=0.052); collaborative>individual in the posttest	Additional benefit
H3: The effect of adaptive feedback depends on the learning mode	All outcome measures	No feedback × learning mode interaction; $F_{1,131}$=0.07; P=.80	Not supported

Discussion on Research Question 1 (Adaptive Compared With Static Feedback)

Our findings show that virtual patients with feedback were beneficial overall for the diagnostic accuracy of medical students. Furthermore, we found that the adaptive type of feedback had positive effects on medical students' learning of conceptual and strategic knowledge in the posttest 1 week after the learning session. Both feedback conditions were associated with increases in conceptual knowledge. For strategic knowledge, scores decreased from pretest to posttest in both conditions, but the decrease was smaller in the adaptive feedback condition.

We interpret the findings such that the adaptive feedback helped students to identify their knowledge gaps and, accordingly, determine which feedback information was particularly relevant to remember. This interpretation is in line with the findings of a study where adaptive feedback was implemented in teacher education [8].

The differential effects observed for diagnostic accuracy, conceptual knowledge, and strategic knowledge may reflect the distinct nature of these constructs. Diagnostic accuracy captures the correctness of a final decision within a specific case, whereas conceptual knowledge reflects an understanding of medical concepts and relationships, and strategic knowledge captures how to approach diagnostic problems. Adaptive feedback may primarily support the development of conceptual understanding and diagnostic strategies rather than directly improving immediate diagnostic accuracy within individual cases.

Recent research on AI-supported learning environments highlights the potential of intelligent systems to support learners' reasoning processes while also raising questions about trust and transparency in AI-assisted feedback [21].

Discussion on Research Question 2 (Individual Compared With Collaborative Learning Mode)

Our results indicate that students in the collaborative learning mode had higher diagnostic accuracy and higher strategic knowledge on the posttest. This pattern of results found in medical students partly deviates from the results found in teacher education (preservice teachers). In teacher education, diagnostic tasks typically involve analyzing classroom situations in order to identify learning difficulties, behavioral issues, or developmental problems in students. These tasks differ from medical diagnosis in that they often allow multiple plausible interpretations and rely more strongly on contextual information. In such situations, collaboration may sometimes increase interpretive divergence rather than convergence, which could explain why collaborative learners in teacher education settings did not always outperform individual learners. More precisely, preservice teachers who worked collaboratively and received static feedback had a significantly lower diagnostic accuracy than learners who worked in the collaborative learning mode and had adaptive feedback. In contrast, medical students who worked in the collaborative learning mode in the present study outperformed students who worked individually, independent of the type of feedback they received, an effect that has been shown before [9]. The study in teacher education assumed negative effects of the unfamiliar collaboration setting [8]. Compared to preservice teachers, medical students might be more used to collaborating because it is a key role of the medical profession [22]. Consequently, they might be more familiar with and less challenged by collaborative learning settings. Besides familiarity with collaborative learning, evidence indicates that students in teacher education show a rather data-driven approach in their diagnostic reasoning, focusing on generating and evaluating evidence as a basis for drawing conclusions; by comparison, students in medical education apply a more hypothesis-driven approach, in which generating and comparing differential diagnoses is particularly relevant [23]. Research has found that generating hypotheses is facilitated in collaborative compared with individual learning settings [11]. Therefore, the benefits of collaboration, such as co-constructing ideas (eg, hypotheses), might rather come into play more in medical education [23]. It seems that the added complexity and students' feedback to each other might not hinder the positive effects of collaboration. Another positive effect of collaboration on identifying relevant differential diagnoses might contribute to an explanation of the higher diagnostic accuracy in the collaborative learning mode and the better retention of strategic knowledge.

Another explanation could be that the diagnoses of classroom situations and virtual patient cases tend to differ in terms of their ambiguity. If a multitude of different perspectives can potentially be taken on a particular case, this may lead to greater divergence (and thus decreased learning outcomes) among collaborative learners—this has been shown, for instance, in studies on computer-supported case-based learning among teachers in adult education [21]. A more detailed analysis of the learning processes will help us to better understand why student teachers and medical students seem to benefit from collaborative diagnosing in different ways.

Interestingly, experts initially disagreed on the solution for one of the virtual patients and reached consensus only after discussion. This observation highlights the inherent complexity of clinical diagnostic reasoning. Even when a case has a correct final diagnosis, different reasoning pathways may lead to that conclusion. Such complexity reflects authentic clinical practice, where diagnostic reasoning often involves weighing competing hypotheses and interpreting incomplete information.

Limitations and Further Research

Generalizability of our findings is limited. We had a limited number of cases from 2 content areas. However, the 2 leading symptoms of back pain and fever that we implemented were chosen to represent relevant fields of clinical practice and cover a wide range of potential diagnoses. Furthermore, we were able to identify small to medium-sized effects with our sample.

Before the study, we had to train an NLP-based system [8,16] in order to provide adaptive feedback: even in the age of quick advances in AI technology, the effort required to implement such a system is still substantial and may not be feasible for smaller medical schools. However, over time, the effort will pay off, and the more cases that are incorporated, the faster it will be to establish an adaptive feedback system. For students, it was a novel experience to receive automated feedback on their written explanations directly in the learning environment, and to understand how the feedback, adapted to their answers, might have influenced the results and provided additional load. To counter this effect, highlighting which part of the feedback response related to which part of the students' written explanations likely increased the effectiveness and acceptance of the adaptive feedback [21]. It should be noted that the present study was a controlled laboratory study—if the described approaches for optimizing learning with virtual patients are actually to be established in medical education, the effects found must also be demonstrated under field conditions. For future use of the system under field conditions, we have high hopes that the adaptive feedback will be widely accepted. The CASUS system, into which we integrated the adaptive feedback [15] and the collaborative learning [16], is widely known to the medical students in the sample.

Although the reliability coefficients of the knowledge tests were moderate, they are comparable to values reported in other exploratory studies using key feature formats. The key feature approach focuses on critical decision points in clinical reasoning rather than broad knowledge coverage, which may result in lower internal consistency but higher construct relevance for diagnostic reasoning tasks.

Additionally, the unequal group sizes between the collaborative and individual conditions may have influenced

statistical power and should therefore be interpreted with caution.

With regard to collaboration, we implemented dyads in our study. It will be interesting to see whether the adaptive form of feedback might benefit larger groups of learners in the system. In future studies, the collaboration itself could also be altered to fully use the potential of the collaborative learning mode, for example, by prompting medical students on how to actively engage in collaboration [22].

From the perspective of clinical reasoning research, collaboration can influence diagnostic performance in several ways [22,23]. Collaborative reasoning may increase the number of diagnostic hypotheses considered, which can reduce the risk of premature closure. At the same time, discussion within a group can introduce cognitive biases such as anchoring or groupthink. The relatively structured nature of the virtual patient cases used in this study may have limited such negative effects, which could explain the positive impact of collaboration observed during the learning phase. To determine the threshold of these negative effects, future studies should introduce increasingly complex virtual patients.

Conclusions

Our adaptive feedback in a virtual patient learning environment provides adaptive elaborated feedback on students' explanations and guidance for improving students' future diagnostic accuracy and the quality of their reasoning and argumentation. Adaptive feedback can advance learning with virtual patients in an individual or collaborative learning mode. The effort it takes to generate adaptive feedback should not be underestimated. However, we argue that despite the effort of developing and training an NLP-based system, this effort might be worthwhile, especially if implementing the adaptive feedback in large programs and offering it to numerous schools or numerous cohorts. As well, future work could explore simplified approaches, such as rule-based feedback or hybrid systems, that could enable broader adoption. Future AI might be able to self-label the texts and minimize the development effort. Until then, an NLP-based system, such as the one presented in this study, seems to be a viable solution for providing a large number of students with elaborated adaptive feedback. It will be interesting to see how large generative AI solutions will be used to create automated adaptive feedback in virtual patient learning environments.

Generative AI Use

During the preparation of this manuscript, the authors used ChatGPT-6.0 (OpenAI) to assist with language editing and refinement. All AI-assisted content was carefully reviewed and revised by the authors, who take full responsibility for the content of the final publication.

Funding

This work was supported by the Federal Ministry of Education and Research (grant 16DHL1039).

Conflicts of Interest

None declared.

References

1. Berman NB, Durning SJ, Fischer MR, Huwendiek S, Triola MM. The role for virtual patients in the future of medical education. Acad Med. Sep 2016;91(9):1217-1222. [doi: 10.1097/ACM.0000000000001146] [Medline: 26959224]
2. Huwendiek S, De leng BA, Zary N, Fischer MR, Ruiz JG, Ellaway R. Towards a typology of virtual patients. Med Teach. Aug 2009;31(8):743-748. [doi: 10.1080/01421590903124708] [Medline: 19811212]
3. Kiesewetter J, Sailer M, Jung VM, et al. Learning clinical reasoning: how virtual patient case format and prior knowledge interact. BMC Med Educ. Mar 14, 2020;20(1):73. [doi: 10.1186/s12909-020-1987-y] [Medline: 32171297]
4. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res. Jul 2, 2019;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
5. Wisniewski B, Zierer K, Hattie J. The power of feedback revisited: a meta-analysis of educational feedback research. Front Psychol. 2019;10:3087. [doi: 10.3389/fpsyg.2019.03087] [Medline: 32038429]
6. Hattie J, Timperley H. The power of feedback. Rev Educ Res. Mar 2007;77(1):81-112. [doi: 10.3102/003465430298487]
7. Chamberland M, Mamede S, St-Onge C, Setrakian J, Bergeron L, Schmidt H. Self-explanation in learning clinical reasoning: the added value of examples and prompts. Med Educ. Feb 2015;49(2):193-202. [doi: 10.1111/medu.12623] [Medline: 25626750]
8. Sailer M, Bauer E, Hofmann R, et al. Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. Learn Instr. Feb 2023;83:101620. [doi: 10.1016/j.learninstruc.2022.101620]
9. Monteiro SD, Sherbino J, Schmidt H, Mamede S, Ilgen J, Norman G. It's the destination: diagnostic accuracy and reasoning. Adv Health Sci Educ Theory Pract. Mar 2020;25(1):19-29. [doi: 10.1007/s10459-019-09903-7] [Medline: 31332589]
10. Kirschner F, Paas F, Kirschner PA. A cognitive load approach to collaborative learning: united brains for complex tasks. Educ Psychol Rev. Mar 2009;21(1):31-42. [doi: 10.1007/s10648-008-9095-2]

11. Hautz WE, Kämmer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. JAMA. Jan 20, 2015;313(3):303-304. [doi: 10.1001/jama.2014.15770] [Medline: 25603003]
12. Vygotsky LS. Cole M, John-Steiner V, Scribner S, Souberman E, editors. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press; 1978. ISBN: 9780674576285
13. CASUS – virtual patient e-learning & online education software [Article in German]. CASUS in.structR. URL: https://www.instruct.eu/casus/ [Accessed 2026-09-25]
14. Bichler S, Sailer M, Bauer E, Fischer F. Fostering diagnostic reasoning in teacher education through authentic learning: the format of simulated cases. OSF. URL: https://osf.io/4gr8t/ [Accessed 2026-09-25]
15. Bauer E, Sailer M, Kiesewetter J, et al. Learning to diagnose students' behavioral, developmental, and learning disorders in a simulation-based learning environment for pre-service teachers. In: Fischer F, Opitz A, editors. Learning to Diagnose with Simulations. Springer; 2022:97-107. [doi: 10.1007/978-3-030-89147-3_8]
16. Kiesewetter J, Hege I, Sailer M, et al. Implementing remote collaboration in a virtual patient platform: usability study. JMIR Med Educ. Jul 28, 2022;8(3):e24306. [doi: 10.2196/24306] [Medline: 35900827]
17. Pfeiffer J, Meyer CM, Schulz C, et al. FAMULUS: interactive annotation and feedback generation for teaching diagnostic reasoning. Proc 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process (EMNLP-IJCNLP): Syst Demonstr. 2019:73-78. [doi: 10.18653/v1/D19-3013]
18. Eclipse INCEpTION. URL: https://inception-project.github.io/ [Accessed 2026-09-25]
19. Schulz C, Sailer M, Kiesewetter J, et al. Fallsimulationen und automatisches adaptives feedback mittels künstlicher intelligenz in digitalen lernumgebungen [Report in German]. Leibniz Institute for Knowledge Media (IWM); 2017. URL: https://www.e-teaching.org/etresources/pdf/erfahrungsbericht_2017_schulz-et-al_fallsimulationen-und-automatisches-adaptives-feedback-mittels-kuenstlicher-intelligenz-in-digitalen-lernumgebungen.pdf [Accessed 2026-09-25]
20. Cohen J. A power primer. Psychol Bull. Jul 1992;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
21. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud. Feb 2021;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
22. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach. Sep 2007;29(7):642-647. [doi: 10.1080/01421590701746983] [Medline: 18236250]
23. Bauer E, Fischer F, Kiesewetter J, et al. Diagnostic activities and diagnostic practices in medical education and teacher education: an interdisciplinary comparison. Front Psychol. 2020;11:562665. [doi: 10.3389/fpsyg.2020.562665] [Medline: 33192833]

Abbreviations

ECG: electrocardiogram
NLP: natural language processing
OSF: Open Science Framework

Edited by Blake Lesselroth; peer-reviewed by Sergio Guinez-Molinos, Stamatis Papadakis; submitted 16.May.2025; initial comments to author 30.Oct.2025; final revised version received 31.Aug.2026; accepted 02.Sep.2026; published 27.Nov.2026

Please cite as:
Kiesewetter J, Werl H, Nederstigt K, Sarbu-Rothsching S, Sailer M, Bauer E, Zottmann JM, Fischer F, Fischer MR
Learning Clinical Reasoning With Virtual Patients Using Adaptive vs Static Feedback in Individual and Collaborative Learning Modes: Randomized Controlled Trial
JMIR Med Educ 2026;12:e77606
URL: https://mededu.jmir.org/2026/1/e77606
doi: 10.2196/77606

© Jan Kiesewetter, Helene Werl, Katharina Nederstigt, Sven Sarbu-Rothsching, Michael Sailer, Elisabeth Bauer, Jan M Zottmann, Frank Fischer, Martin R Fischer. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.