Fostering Diagnostic Decision-Making in Special Education Students through Simulation-Based Games and AI Pedagogical Agents: A Case-Based Learning Comparison in Higher Education

Judith Zellner[1], Jakob Koch[1], Maximilian Fink[2], Nikola Ebenbeck[1], Markus Gebhardt[1]

[1] Ludwig-Maximilians-Universität München, Germany
[2] Bundeswehr-Universität München, Germany

HOW TO CITE:
Zellner J., Koch J., Fink M., Ebenbeck N., Gebhardt M. (2025). Fostering Diagnostic Decision-Making in Special Education Students through Simulation-Based Games and AI Pedagogical Agents: A Case-Based Learning Comparison in Higher Education. *International Journal of Special Education, 40*(2), 92-103.

CORRESPONDING AUTHOR:
Judith Zellner;
judith.zellner@edu.lmu.de

DOI:
https://doi.org/10.52291/ijse.2025.40.24

ABSTRACT:

Educational diagnostic skills are essential in special education, both for identifying students' special needs and for making informed pedagogical decisions in daily practice. Simulation-based learning presents a promising approach in higher education by connecting theoretical knowledge to real-world problems and enabling in-depth analysis of diagnostic processes through the generation of data. This study examines the use of simulations in diagnostic case processing, comparing two formats: a digital structured click game and a discursive interaction with an AI-based Pedagogical Agent (PA). Both games were developed for students with special education needs who possess expertise in diagnostic processes. Findings indicate that both interactive learning environments effectively support the development of diagnostic skills and processes. All participants successfully completed the games, although 75% required additional support during the AI conversation. Participants demonstrated significantly varying levels of efficiency and accuracy in their diagnostic process across the two games. This paper explores the underlying reasons for these differences and discusses the potential benefits and limitations of interactive learning environments, with and without AI integration.

Keywords: Diagnostic Decision-Making, Teacher Education, AI-based Agents, Simulation-based Learning, Personalized Learning

INTRODUCTION

AI-based Pedagogical Agents in higher education

AI-based PAs are advanced, interactive tools designed to enhance teaching and learning experiences. Typically embodied as 3D models, these agents facilitate real-time interactions through conversation and other modalities, powered by artificial intelligence. This enables them to adapt to learners' needs dynamically, providing personalized and immersive educational experiences (Zhao et al., 2024). In this paper, we employ an AI-based PA to support the training of diagnostic decision-making (DDM) for special education students. By interacting with the PA, students practice gathering and evaluating information, making educational decisions, and justifying their reasoning in a realistic, simulated environment.

PAs offer a wide range of benefits for teaching and learning, particularly in areas such as individual adaptation, contextualized learning, immersive learning, scaffolding, and promoting self-regulation and motivation (Fink et al., 2024). They enable flexible adaptation to learners' needs, ensuring that tasks are aligned with individual levels and progress (Wollny et al., 2021). Such personalization enhances motivation and engagement (Shumanov & Johnson, 2021) and allows for the efficient creation of contextualized scenarios. In medical and teacher education, avatar-based role plays foster communicative skills and complement or replace traditional formats (Fecke et al., 2023). Research underscores the added value of PAs in supporting learning and motivation (Kim, 2009; Siegle et al., 2023). They also contribute to immersive virtual and augmented environments by preventing cognitive overload and integrating established learning strategies (Makransky et al., 2019). Through scaffolding and tailored feedback, they facilitate better outcomes across subjects (Sailer et al., 2023). Moreover, they foster self-regulation, positive emotions, and curiosity, which are central to learning success (Beege & Schneider, 2023).

However, these benefits are accompanied by critical challenges. Simulation-based learning can overwhelm learners, who must apply their knowledge in complex professional situations (Fischer et al., 2022; Frerejean et al., 2023; Machts et al., 2024). Adequate instructional support, such as targeted feedback, is therefore essential (Issenberg et al., 2005). While personalization has been studied in specific contexts, systematic integration into simulation-based formats for professional learning remains limited. This gap is particularly relevant because professional development requires practice representations that both approximate real-world complexity and support the structured application of knowledge (Boshuizen et al., 2020; Jossberger et al., 2022; Norman et al., 2007). Against this background, the question arises how training culture can be designed to foster educational diagnostic competences.

Training culture to foster educational diagnostic competencies

Teachers' diagnostic skills are crucial for supporting students' individual learning progress. Research highlights that strong diagnostic skills enhance the likelihood of successful learning outcomes in the classroom. This is achieved by adapting lessons to meet students' individual learning needs (Blömeke et al., 2008). In recent years, the importance of acquiring these diagnostic skills has grown across all teacher training and professional development programs, reflecting its critical role in effective teaching (e.g., Südkamp et al., 2012).

Decision-making in the diagnostic process

The process of educational diagnostics involves multiple steps, with varying emphasis depending on the underlying model (e.g., Bundschuh & Winkler, 2019; Gebhardt, 2024). Diagnostic assessment typically begins with an initial concern, followed by the collection of information, the interpretation of results, and ultimately a decision or judgment regarding support needs (van Ophuysen & Behrmann, 2015). Data collection is guided by the diagnostic purpose and the attributes being measured (Tönnissen & Hövel, 2022). Effective diagnostics rely on the selection of suitable methods to provide relevant and comprehensive information, rather than intuitive predictions (van Ophuysen & Behrmann, 2015).

Educational diagnostics is a cyclical process that integrates diagnostic and pedagogical decisions to personalize learning support (Gebhardt, 2024). This cycle involves making support decisions, implementing targeted strategies, and evaluating learning progress (Heitzmann et al., 2019). While flexibility, efficiency, and the ability to revise decisions enhance the diagnostic process, these elements remain underexplored in complex tasks requiring advanced information processing (Dünnebier et al., 2009; Bauer et al., 2025). Interpreting results to prioritize relevant aspects and inform support decisions requires situating insights within theoretical and contextual frameworks (Tönnissen & Hövel, 2022). Appropriate support decisions integrate diagnostic interpretations with pedagogical expertise, enabling the implementation

of evidence-based strategies (Hövel et al., 2019). Effective decisions can often be made with limited empirical evidence (Reimer et al., 2007), striking a balance between accuracy and efficiency to ensure responsive, developmentally appropriate support.

Fostering diagnostic competencies

The framework proposed by Heitzmann et al. (2019) systematically examines the development of diagnostic competencies, identifying core quality criteria that influence the diagnostic process. These include diagnostic knowledge, activities, and context. Diagnostic knowledge encompasses conceptual knowledge, such as theoretical understanding, and strategic knowledge, including heuristics and procedures (Förtsch et al., 2018; Stark et al., 2011). Together, these knowledge types support informed diagnostic decisions. Accuracy, which measures alignment with expert solutions, and efficiency, referring to the resources required, are key indicators of diagnostic quality (Heitzmann et al., 2019).

Enhancing diagnostic competencies involves the transfer of theoretical knowledge and application-based learning. While lectures and self-study focus on knowledge dissemination, applying skills often includes interactive or collaborative case analyses (Heitzmann et al., 2019). However, university teacher education is often criticized for insufficient preparation in real-world application (Harr et al., 2014). Bridging theory and practice requires opportunities that combine theoretical foundations with authentic experiences, enabling students to develop adaptive responses and sustainable routines (Renkl, 2014; Seidel et al., 2015).

Problem-based learning fosters a connection between theoretical knowledge and practical application, emphasizing key activities such as problem identification and evidence gathering (Fischer et al., 2014). Its effectiveness depends on instructional support, such as scaffolding strategies and prompts, but these must be carefully timed (Renkl, 2014; Kim et al., 2018). Case vignettes, including written or video-based analyses, promote foundational skills such as identifying relevant aspects; however, their limited interactivity constrains their ability to replicate real teaching scenarios. Interactive simulations address this gap by offering controlled, dynamic environments for skill practice and reflection on mistakes (Cook, 2014). These methods, however, require customization to diverse learner needs (Chernikova et al., 2020), and further research is needed to explore their effectiveness across diverse contexts (Cook, 2014).

RESEARCH QUESTIONS

Developing DDM skills is crucial for preparing future educators to confidently navigate real-world, learner-centered contexts. Simulated environments that replicate diagnostic challenges play a critical role in bridging the gap between theoretical knowledge and practical application. However, this gap often persists, underscoring the need for more robust training methods. In this study, students at the university level engage with two distinct tools, which we compare in this study: a click-based simulation game (click game) and an AI-based PA (AI games). Through this comparative analysis, we aim to explore how each tool supports the development of diagnostic proficiency and heuristics as complementary resources for diagnostic training. Therefore, we address two research questions:

(A) How do the games differ in their effectiveness for training teachers' diagnostic competencies, and what types of support measures are necessary to ensure their successful implementation in higher education?

(B) Do students demonstrate distinct solution strategies and deeper reflection when engaging with the diagnostic process in these games?

METHODS

Sample and prerequisites for participation

The study involved N = 8 students (four male, four female) enrolled in a special education teacher training program at a German university, with a mean age of M = 22 years (SD = 1.6). At the time of data collection, students were in their 4th semester, engaged in coursework on educational diagnostics. This included a lecture and a practice seminar focused on test theory and diagnostic procedures, where students practiced applying theoretical content on test theory, diagnostic procedures, and data-based DDM using document-based case vignettes to strengthen their applied understanding.

Before the study, students assessed their diagnostic competencies using the validated DaKI self-assessment questionnaire (Jungjohann & Gebhardt, 2023). At the outset, estimating the required effort and potential costs was challenging, as this study was among the first to implement the avatar format in teaching with university students. We assumed that substantial prior knowledge would be necessary to successfully complete the tasks. Selection criteria emphasized self-rated proficiency and confidence in educational diagnostics, as we hypothesized

that a positive self-concept and foundational diagnostic heuristics would be essential for success in the interactive learning environment (Zellner et al., 2024). Only students with an above-average self-assessment score were included in the study (M = 4.31, SD = 0.18). All participants engaged in two tasks: completing one case in the click game and another in the AI game.

Instruments

Both the click game and the AI game were designed to follow the same diagnostic process, which is structured into four steps (based on, e.g., Bundschuh & Winkler, 2019). However, the games differ in their level of structure.

1. **Step 1 - Initial Information Gathering:** Students begin by collecting foundational information about the case. This involves formulating a problem, conducting assessments, and developing initial hypotheses.
2. **Step 2 - Systematic Verification and Supplementation:** Students systematically verify and expand upon the information gathered in Step 1 using a combination of formal and informal diagnostic methods.
3. **Step 3 - Synthesis and DDM:** Students focus on synthesizing the collected information into a categorical decision.
4. **Step 4 - Intervention Planning:** Students formulate specific intervention measures grounded in the diagnostic results and their interpretations (Hövel et al., 2019).

The case vignettes used in the two games present a student with reading difficulties. Using appropriate diagnostics, the participants can determine that the student has difficulty with the phoneme-grapheme correspondence of certain sound combinations. This needs to be automated so that students reach the next stage of development and achieve fluent reading.

Click-based simulation game

The click game (Figure 1A) is a text-based learning format designed to guide students through the DDM process in a structured way. It is technically implemented with the Socisurvey software (Leiner, 2024). Students first receive a written introduction to the diagnostic process requirements within the game. They are then presented with a fictional case, followed by structured instructions and DDM questions. Responses to these questions are selected by clicking, allowing for an interactive learning experience.

In the case scenario, a simulated teacher provides a detailed case description that includes the student's learning behaviors, specific difficulties, and additional context, such as potential distractions (e.g., motor issues). After reviewing all the provided information, participants are instructed to select an appropriate diagnostic test from a list of five options. To assist with their decision, students have the option to review an additional support text summarizing key theoretical concepts in standardized testing. Of the five diagnostic tests offered, only one is directly relevant to the learning difficulty described in the case, while the others are not aligned with the identified difficulty. The click game concludes with two open-ended questions that prompt participants to interpret the test results and propose suitable support measures for the student.

AI-based Pedagogical Agent

For the AI game (Figure 1B), GPTAvatar (Robinson, 2023), an AI-based PA, designed to provide a realistic, interactive learning experience, is utilized. The software integrates automatic speech recognition to convert spoken input into text, which is then processed by a Large Language Model (LLM), such as OpenAI GPT-4, to

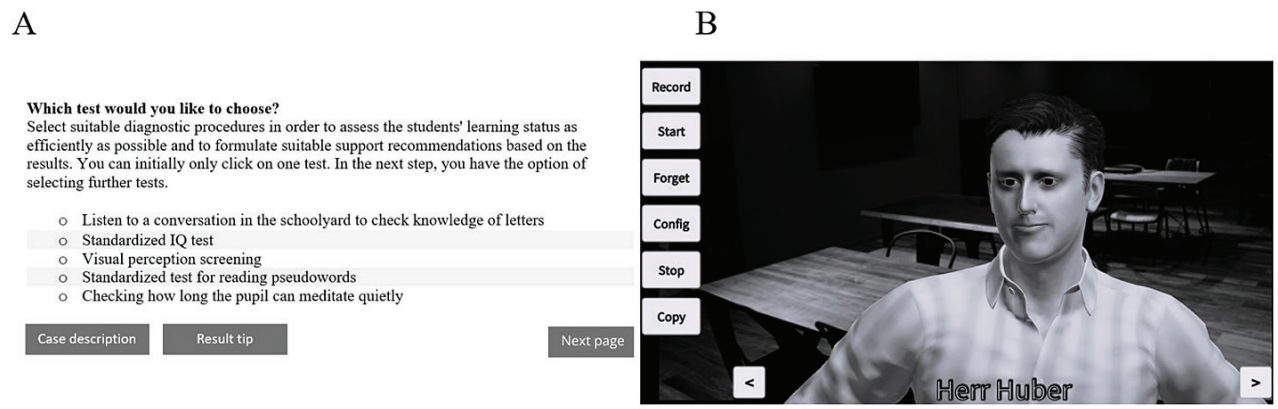

Figure 1. Interactive learning environments: (A) click-based simulation game and (B) AI-generated PA

Note: Figure 1A has been translated into English for better comprehension. The original click game is in the German language.

generate responses. Text-to-speech technology outputs these responses in natural-sounding synthetic speech. Realistic animations and lip synchronization are powered by the Unity Engine and SALSA. Data collection utilized a desktop version of GPTAvatar (Robinson, 2023), with an extension that saved log data, including full dialogues with timestamps (Fink, 2024).

The modular architecture enables customization of the PA's personality, scenario, and responses via a config file, allowing personalized learning environments. In this case, participants engaged in a simulated collegial consultation with the PA, portraying "Mr. Huber," a fictional teacher discussing a struggling student. The PA followed a pre-programmed case vignette and structured diagnostic steps identical to the click game, ensuring comparability. Strict rules defined the PA's behavior, requiring adherence to the diagnostic process, providing hints without directly offering solutions, and maintaining a fully verbal interaction. Conversations began with the standardized prompt: *"Hello, I have a student who is struggling with reading. I am wondering how to support her appropriately. Would you have a moment to discuss the case with me?"* After this, interactions unfolded dynamically based on participant input.

Participants received optional support, including a brief orientation by a trained facilitator, a guidance document outlining the scenario and objectives, and the option to initiate two practice exchanges with the PA before the main session. The main session, set in an inclusive school context, began with a reset and the PA's introductory interaction. If participants struggled, a flowchart summarizing the four diagnostic steps and their goals was provided to guide them.

Questionnaire

To evaluate students' subjective experience regarding their motivation, the usability of the games, and their perceived diagnostic skills, we employed a digital self-assessment questionnaire implemented in Socisurvey (Leiner, 2024). The questionnaire utilized a five-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5). The motivation scale was adapted from Bandura & Schunk (1981), the diagnostic skill scale was based on the DaKI questionnaire by Jungjohann & Gebhardt (2023), and the usability items were derived from the TPACK model by Schmid & Petko (2020). Additionally, students were invited to provide qualitative feedback through two open-ended questions, allowing them to highlight particularly positive aspects and suggest areas for improvement.

Setting

Data collection was conducted as part of a practical seminar on educational diagnostics during the final weeks of the summer semester in 2024. The click game was administered two weeks before the AI game. Students completed the click game on their personal mobile devices. The AI game took place individually in a quiet room equipped with an internet-enabled laptop, speakers, and a microphone. After completing the AI game, students filled out the self-assessment questionnaire.

Analysis

All analysis was carried out in R and RStudio (R Core Team, 2018). First, a descriptive analysis was conducted to examine overall interaction with the digital learning environments. The second step focused on students' behavior while engaging with the games. This analysis included quantitative variables derived from the framework by Heitzmann et al. (2019), such as time spent on the games, activity levels within the games, the efficiency and accuracy of diagnostic decisions, and the use of support measures. Qualitative data related to students' interpretations and proposed support measures were analyzed using a content-analytic approach. Based on predefined categories, the responses were coded as insufficient, sufficient but not data-driven, and sufficient and data-driven. Coding was carried out independently by two evaluators. Any disagreements were discussed until a consensus was reached.

RESULTS

Results of interactive games

The two games differ in the level of structuring they provide for the learning process, and students varied in terms of time spent, the activity level, and the efficiency and accuracy of their diagnostic decisions.

Duration of Interactive Games

Participants required significantly more time to complete the click game than to work with the AI game (Figure 2). On average, the click game took M = 10.02 minutes (SD = 5.26, Min = 2.05, Max = 17.1). In contrast, the AI game lasted M = 4.99 minutes (SD = 1.97, Min = 2.87, Max = 7.93). This difference in duration was statistically significant ($p < .05$) with a very large effect (Cohen's D: d = 1.27). Individual students displayed variation in game duration of up to 304 seconds (approximately 5 minutes). The proportion of verbal exchanges between students and the AI-based PA was evenly distributed, with

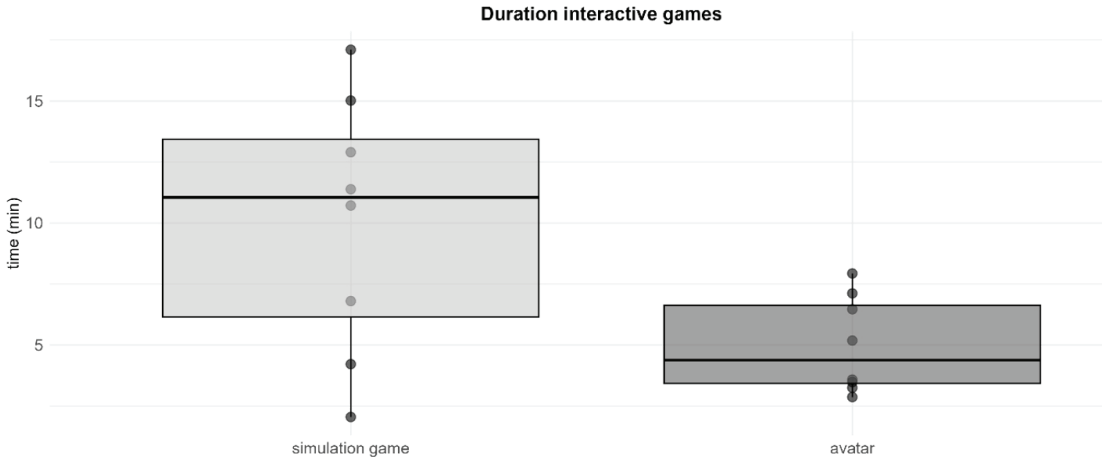

Figure 2. Duration of the interactive games.

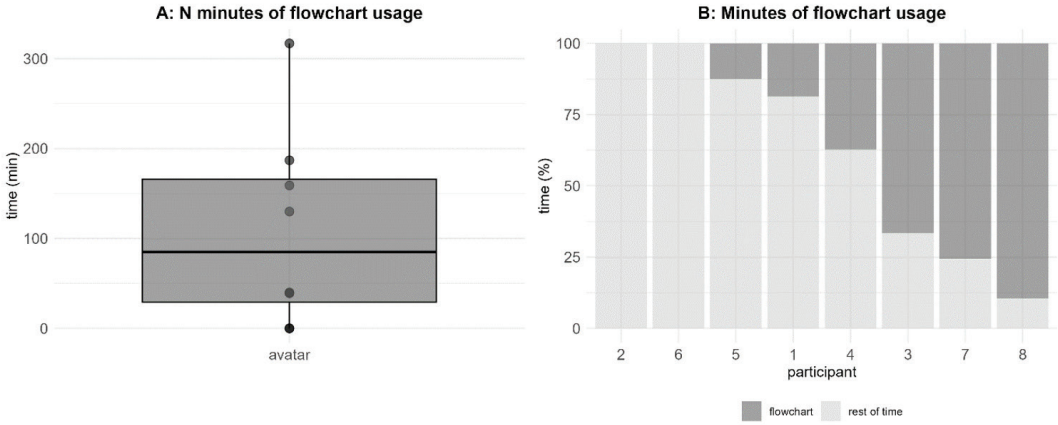

Figure 3. Usage of the flowchart.

a consistent 50:50 ratio across all steps. Steps 1 and 2 exhibited the highest levels of interaction, which aligned with the observed time spent on these steps. Conversely, Step 3 (interpretation) showed the lowest interaction overall. (Figure 2).

Instructional support

Most students relied on the provided instructional aids and scaffolding tools, such as the flowchart, to successfully navigate the games and identify appropriate support measures. In the click game, 87.5% of students accessed additional instructional materials on standardized tests. In the AI game, 75% of students utilized the flowchart as a supplementary guide and scaffolding tool. As shown in Figure 3, students accessed the flowchart on average *M* = 2.42 minutes after starting the AI game (*SD* = 1.73, *Min* = 0.65, *Max* = 4.63). The flowchart was consistently used as a transition aid between steps. Additionally, all students used the optional support provided by the AI-based PA at the end of each step. While some students used the PA's suggestions to independently formulate their conclusions, others directly sought solutions by asking the AI questions, such as, *"What would you do?"* (Figure 3).

Quality of diagnostic decisions

Students demonstrated varying levels of efficiency and accuracy in their DDM across the two games. In the click game, all students successfully identified the correct diagnostic tool. On average, students selected *M* = 3.62 tests (*SD* = 1.06), resulting in a hit rate of *M* = 30% (*SD* = 9.88%). In the AI game, students selected fewer tests on average (*M* = 1.25, *SD* = 0.46). The majority of students (75%) identified the correct test; one student achieved a hit rate of 50%, and one student failed to select a correct test. The overall hit rate for the AI game was M = 81.25% (SD = 37.2%). A significant difference in hit rates was observed between the two games ($p < .05$), with a very large effect size (Cohen's D: d = 1.88). (Figure 4).

Students' proposed support measures and interpretations were categorized as either insufficient, sufficient, and not data-based, or *sufficient and data-based*. Students

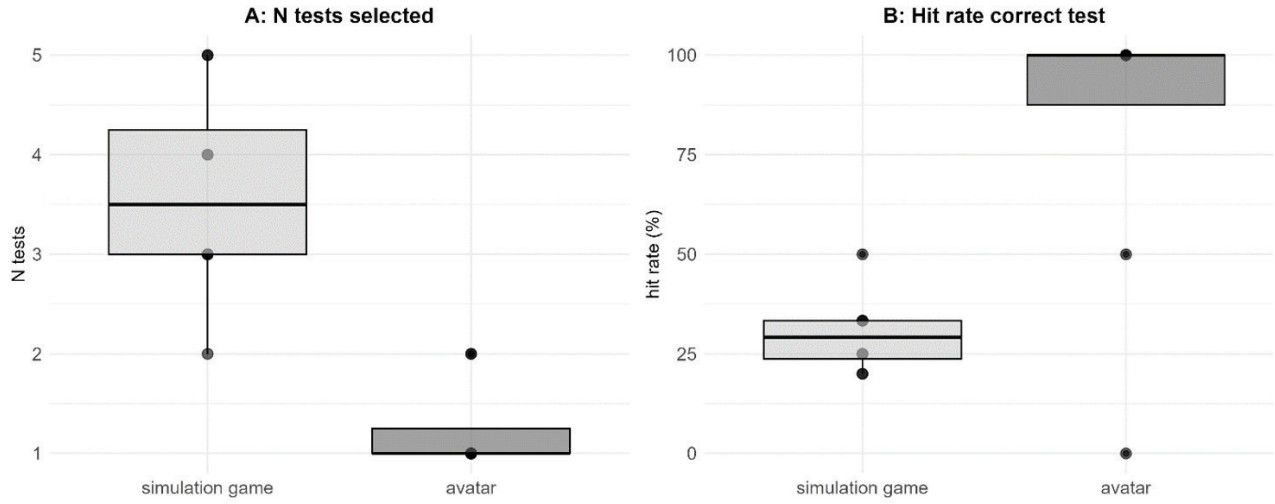

Figure 4. N and proportional hit rate.

with sufficient and data-based interpretations responded e.g.,: *"L. is at the late alphabetic level. She links letters to sounds and sounds words correctly, but has difficulties with reading fluency and decoding more complex sound combinations such as /ie/. This makes the transition to the orthographic level more difficult for her"*. In the click game, 37.5% of students achieved data-based and sufficient interpretations, compared to 62.5% in the AI games. Responses classified as sufficient but not data-driven (e.g., "*The girl shows an abnormality in letter awareness*") accounted for 50% in the click game and 25% in the AI game. Insufficient interpretations (e.g., "*Promotion of perceptual competence*") were noted in 12.5% of cases in both games.

Students with data-based and sufficient support measures answered, e.g., *"1. Exercises with frequent sound combinations and syllable structures to improve decoding skills. 2. Repeated reading of simple texts to increase reading speed 3. Discussion of simple text questions to promote text comprehension"*. In the click game, 37.5% of students identified sufficient and data-based support measures, compared to 75% in the AI game. All other students in both games provided responses that were sufficient but not data-driven, and in some cases, far removed from the case, e.g., "*Syllable games, rhymes, primers with syllable sheets, larger font, hearing/recognizing vowels*".

When analyzing the participants' overall interaction with the games, significant overlap was observed in the diagnostic pathways of most students (75%). Steps 1 (initial problem formulation) and 4 (support measures) showed the greatest homogeneity in responses. Steps 2 (information collection) and 3 (interpretation) displayed greater variability, with some students achieving notably higher or lower hit rates and data-based sufficiency in their interpretations and support measures. In line with Research Question B, the results show that students demonstrated deeper reflection in the AI game, as evidenced by the higher proportion of data-based and sufficient interpretations and support measures compared to the click game.

Self-assessment of the subjective experience of interaction with an AI-based PA

The questionnaire results demonstrated high internal consistency across all dimensions, with a Cronbach's alpha of 0.93. The diagnostic skill dimension yielded an average score of 4.06 *(SD* = 0.76). The motivation dimension scored the highest, with an average of 4.59 (*SD* = 0.67). The technology dimension received a mean score of 4.1 (*SD* = 0.93). Open-ended responses highlighted several perceived benefits of the AI game, including the opportunity for direct interaction with the PA (1 student), the ability to ask questions without requiring extensive prior knowledge (6 students), individualized responses (2 students), and the structured flowchart provided (1 student). Students also suggested areas for improvement, including clearer guidance and examples of potential interactions with the PA (6 students), as well as a dedicated space for recording information during the activity (1 student). By this, participants referred to short prompts illustrating possible student–agent dialogues, such as asking for clarification, proposing a test, or requesting additional case information.

DISCUSSION

The click game and AI game on DDM, as described in this study, effectively support the development of diagnostic skills and strategic knowledge, particularly

in applying heuristics and structured diagnostic procedures (Förtsch et al., 2018; Stark et al., 2011) within a university context. Feedback from students who tested both systems indicated a high level of satisfaction, aligning with previous research showing increased motivation for interactive and AI-learning environments (e.g., Kim, 2009). The implementation of both games exceeded expectations from both pedagogical and technical perspectives.

Quality of diagnostic decisions

Despite their prior confidence in diagnostic abilities (M = 4.31, SD = 0.18) and foundational training in diagnostic basics, students initially demonstrated lower-than-expected efficiency and accuracy in DDM. The quality of decisions varied significantly across processes and tools, reflecting a gap between theoretical knowledge and practical application, particularly in problem recognition and DDM quality (Fischer et al., 2014).

Significant variation was observed in the choice of diagnostic tools, the quality of interpretation, and overall diagnostic outcomes among participants. In the click game, no student achieved 100% accuracy in test selection, with an average hit rate of 30%. Conversely, in the AI game, six students achieved perfect scores, leading to a higher mean hit rate of 81%. These findings suggest that the type of learning tool influences diagnostic performance, although a sequence learning effect cannot be ruled out. Participants who performed poorly in the click game often showed improvement in the AI game, while one high-performing click game participant struggled with the less structured and faster-paced AI game environment. Students' familiarity with the click game's format may encourage experimentation, resulting in lower hit rates, while the AI game's novel methodology and interactive dialogue may prompt more cautious, accuracy-focused behavior.

Students generally excelled at problem-solving but struggled with recognizing and accurately identifying problems. Data-driven, case-appropriate interpretations were more common in the AI game, whereas the click game often yielded sufficient but less evidence-based interpretations. These challenges underscore the difficulty of integrating theoretical knowledge with diagnostic data and prioritizing interventions when multiple issues arise (Tönnissen & Hövel, 2022). A lack of evidence-based DDM compromises support measures, which must adhere to high-quality standards (Gebhardt, 2024).

Structuring and Scaffolding in the Learning Environments

Students required structured guidance in the AI game, particularly when transitioning between steps in the diagnostic process. The flowchart was frequently used as a transitional aid, particularly when transitioning from test results to data-driven interpretations. Many appreciated the ability to ask questions and use the flowchart, which resulted in increased confidence and clarity. This aligns with research showing scaffolding enhances autonomy, albeit with varying effects (Heitzmann et al., 2018; Stark et al., 2011). Notably, two students navigated the diagnostic process without the flowchart, indicating that prior training may have enabled them to internalize the steps. However, the variation raises the question of why some students were able to work independently while others depended on support. It is also unclear whether students used the flowchart out of necessity or as a safeguard due to insecurity.

Efficiency of the games

The AI game resulted in more efficient and accurate diagnostic decisions in less time compared to the click game. Students spent significantly less time in the AI game (M = 4.99 minutes, SD = 1.97) than in the click game (M = 23.12 minutes, SD = 13.57), despite the additional time required for verbal exchanges with the AI and its integration with advanced tools such as ChatGPT and speech recognition. Verbal case handling in the AI game was faster and more effective, aligning with findings that diagnosing virtual actors is quicker than menu-based cases (Fink et al., 2021).

Students noted that the fleeting nature of AI interactions sometimes hindered reflection, underscoring the need for tools to facilitate revisiting conversations. However, the results suggest that shorter engagement does not compromise learning outcomes. Although the click game required more time, the AI game led to higher-quality interpretations and support measures, with 75% of participants formulating data-driven, case-appropriate measures compared to 25% in the click game. This supports Reimer et al.'s (2007) conclusion that reliable decisions depend more on the quality of evidence than its quantity. The difference lies in the approach: the click game is pre-structured and knowledge-based, encouraging participants to think before acting, whereas the AI game fosters intuitive, self-regulated engagement. By prompting participants to verbalize their thoughts and interact with the AI, the game reduces the cognitive load associated with independent problem-solving. Consequently, the AI game

emphasizes the entire diagnostic process, including its interrelationships, more holistically than document-based case vignettes (Heitzmann et al., 2019; Kim, 2009).

LIMITATIONS

Although reality and student learning are more complex than any simulation can fully capture, students exhibited similar patterns of behavior across the games. Simulations are often critiqued for oversimplifying the complexities of real-world situations and the multifaceted nature of student learning (e.g., Südkamp et al., 2012). Furthermore, for computational reasons, this study categorized responses as either correct (1) or incorrect (0). In reality, responses often fall within a spectrum, being neither wholly correct nor entirely incorrect. Another limitation concerns the small sample size ($N = 8$). The experimental design of the study justifies the reduced number of participants but also restricts the generalizability of findings. Results should therefore be interpreted as indicative rather than representative. There may also be a sequential effect, as all participants played the click game first and then the AI-based game. This order may have influenced the outcomes, for example, by fostering a deeper understanding of the diagnostic process through the initial task, which could in turn enhance subsequent performance in the AI-based game. Future studies will randomize the order of gameplay. Finally, this study does not address the long-term impact of such training on students' diagnostic competencies. The results provide initial insights into short-term processes; however, longitudinal research is necessary to determine whether the observed learning effects persist over time and are effectively transferred into practice.

CONCLUSION AND FURTHER RESEARCH

In summary, all participants successfully engaged in both learning environments, confirming the feasibility of the approach and its potential for innovation in teaching and assessment. We anticipated similar performance in the pre-structured click game and more variable outcomes in the open AI game. However, the findings indicate that differences between the two games emerged mainly in the time taken to complete tasks. Nevertheless, diagnostic quality differed between the two tools, too, with higher accuracy in the AI game than in the click game. This suggests that prior knowledge and established action templates supported students in making structured and accurate decisions, even in more complex scenarios. These results highlight the value of integrating traditional learning strategies into immersive environments, which can deepen learning and enhance success (Makransky et al., 2019). The structured progression of the exercises appears to have contributed to their accessibility and effectiveness, as all participants entered with a high level of prior knowledge.

Since only students with high self-assessed diagnostic competence participated in this study, the variation within this group is noteworthy: 75% required additional support during the AI game in the form of written guidance. This indicates that performance depends not only on the learning tool but also on individual learner characteristics. Whether these differences stem from cognitive load, varying degrees of technical familiarity, or other factors should be examined through systematic follow-up studies.

The AI-based PA performed effectively, primarily due to the rigorous engineering of prompts. One minor technical issue arose with the speech recognition software when a participant spoke in a dialect; however, this is likely to have had no significant impact on overall performance. Ongoing improvements in such technologies should further mitigate these challenges. Both the click game and AI game demonstrate potential as educational tools, training resources, and methods for quantitative research. They offer resource-efficient solutions for collecting and analyzing diagnostic applications while creating engaging and interactive teaching-learning environments. Future research should explore the potential of these interactive learning environments, particularly the AI-based PA, with larger and more diverse samples.

ACKNOWLEDGEMENT

None

DECLARATION OF INTEREST STATEMENT

The authors reported no potential conflict of interest

ETHICAL STATEMENT

This study was conducted in accordance with the principles of ethical research. Prior to data collection, all participants were provided with detailed information about the study's purpose and procedures, and informed consent was obtained before their participation. To ensure confidentiality, all data were anonymized, and participants were assigned codes (e.g., Participant 1, Participant 2) in the reporting of results. All collected data were securely stored. The participants engaged voluntarily in the study and were informed of their right to withdraw from the study at any time without consequence.

FUNDING

None

REFERENCES

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41*(3), 586–598. https://doi.org/10.1037/0022-3514.41.3.586

Bauer, E., Heitzmann, N., Bannert, M., Chernikova, O., Fischer, M. R., Frenzel, A. C., Gartmeier, M., Hofer, S. I., Holzberger, D., Kasneci, E., Koenen, J., Kosel, C., Küchemann, S., Kuhn, J., Michaeli, T., Neuhaus, B. J., Niklas, F., Obersteiner, A., Pfeffer, J., Sailer, M., Schmidmaier, R., Schmidt-Hertha, B., Stadler, M., Ufer, S., Vorholzer, A., Seidel, T., & Fischer, F. (2025). Personalizing simulation-based learning in higher education. *Learning and Individual Differences, 122,* 102746. https://doi.org/10.1016/j.lindif.2025.102746

Beege, M., & Schneider, S. (2023). Emotional design of pedagogical agents: The influence of enthusiasm and model–observer similarity. *Educational Technology Research and Development, 71*(3), 859–880. https://doi.org/10.1007/s11423-023-10213-4

Blömeke, S., Felbrich, A., Müller, C., Kaiser, G., & Lehmann, R. (2008). Effectiveness of teacher education: State of research, measurement issues and consequences for future studies. *ZDM – Mathematics Education, 40*(5), 719–734. https://doi.org/10.1007/s11858-008-0096-x

Boshuizen, H. P. A., Gruber, H., & Strasser, J. (2020). Knowledge restructuring through case processing: The key to generalise expertise development theory across domains? *Educational Research Review, 29*, 100310. https://doi.org/10.1016/j.edurev.2020.100310

Bundschuh, K., & Winkler, C. (2019). *Einführung in die sonderpädagogische Diagnostik* [Introduction to special education diagnostics]. UTB. https://doi.org/10.36198/9783838552866

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., Fischer, F., & DFG Research group COSIMA. (2020). Facilitating diagnostic competences in higher education—a meta-analysis in medical and teacher education. *Educational Psychology Review*, *32*(1), 157-196. https://doi.org/10.1007/s10648-019-09492-2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Cook, D. A. (2014). How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. *Medical Education, 48*(8), 750–760. ttps://doi.org/10.1111/medu.12473

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dünnebier, K., Gräsel, C., & KrolakSchwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung: Eine experimentelle Studie zu Ankereffekten [Judgment biases in school performance assessment: An experimental study on anchoring effects]. *Zeitschrift für Pädagogische Psychologie, 23*(3–4), 187–195. https://doi.org/10.1024/1010-0652.23.34.187

Fecke, J., Afzal, E., & Braun, E. (2023). A conceptual system design for teacher education: Role-play simulations to train communicative action with AI agents. In J. D. Slotta & E. S. Charles (Eds.), *Proceedings of the third annual meeting of the International Society of the Learning Sciences (ISLS)* (pp. 2197–2200). International Society of the Learning Sciences. https://doi.org/10.22318/icls2023.197799

Fink, M. C., Reitmeier, V., Stadler, M., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Assessment of diagnostic competences with standardized patients versus virtual patients: Experimental study in the context of history taking. *Journal of Medical Internet Research, 23*(3), e21196. https://doi.org/10.2196/21196

Fink, M., Robinson, S., & Ertl, B. (2024). AI-based avatars are changing the way we learn and teach: Benefits and challenges. *Frontiers in Education, 9*, 1416307. https://doi.org/10.3389/feduc.2024.1416307

Fischer, F., Chernikova, O., & Opitz, A. (2022). Learning to diagnose with simulations: Introduction. In F. Fischer & A. Opitz (Eds.), *Learning to diagnose with simulations: Examples from teacher education and medical education* (pp. 1–8). Springer International Publishing. https://doi.org/10.1007/978-3-030-89147-3_1

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(2), 28–45. https://doi.org/10.14786/flr.v2i2.96

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M. R., Girwidz, R., Obersteiner, A., Reiss, K., Stürmer, K., Siebeck, M., Schmidmaier, R., Seidel, T., Ufer, S., Wecker, C., & Neuhaus, B. J. (2018). Systematizing Professional Knowledge of Medical Doctors and Teachers: Development of an Interdisciplinary Framework in the Context of Diagnostic Competences. *Education Sciences, 8*(4), 207. https://doi.org/10.3390/educsci8040207

Frerejean, J., van Merriënboer, J.J.G., Condron, C. *et al.* (2023). Critical design choices in healthcare simulation education: a 4C/ID perspective on design that leads to transfer. *Advances in Simulation,* 8, 5. https://doi.org/10.1186/s41077-023-00242-7

Gebhardt, M. (2024). Pädagogische Diagnostik: Leistung, Kompetenz und Entwicklung messen, bewerten und für individuelle Förderung interpretieren [Educational diagnostics: Measuring, assessing, and interpreting performance, competence, and development for individual support]. Munich: LMU Munich. https://doi.org/10.5282/ubm/epub.110013

Harr, N., Eichler, A., & Renkl, A. (2014). Integrating pedagogical content knowledge and pedagogical/psychological knowledge in mathematics. *Frontiers in Psychology, 5*, 924. https://doi.org/10.3389/fpsyg.2014.00924

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda. *Frontline Learning Research, 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Hövel, D. C., Hennemann, T., & Rietz, C. (2019). Meta-Analyse: programmatischer-präventiver Förderung der emotionalen und sozialen Entwicklung in der Primarstufe [Meta-analysis: Programmatic-preventive promotion of emotional and social development in primary school]. *Emotionale und soziale Entwicklung in der Pädagogik der Erziehungshilfe und bei Verhaltensstörungen: ESE, 1*(1), 38–55. https://doi.org/10.25656/01:25182

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher, 27*(1), 10–28. https://doi.org/10.1080/01421590500046924

Jossberger, H., Breckwoldt, J., & Gruber, H. (2022). Promoting expertise through simulation (PETS): A conceptual framework. *Learning and Instruction, 82*, 101686. https://doi.org/10.1016/j.learninstruc.2022.101686

Jungjohann, J., & Gebhardt, M. (2023). Fragebogen zur Erfassung der diagnostischen Kompetenz von Lehrkräften in der inklusiven Schule (DaKI, Version 0.2) [Questionnaire for assessing teachers' diagnostic competence in inclusive schools] [Research instrument]. https://doi.org/10.5283/epub.54249

Kim, N. J., Belland, B. R., & Walker, A. E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for STEM education: Bayesian meta-analysis. *Educational Psychology Review, 30*(2), 397–429. https://doi.org/10.1007/s10648-017-9419-1

Kim, Y. (2009). The role of learner attributes and affect determining the impact of agent presence. *International Journal of Learning Technology, 4*(3–4), 234–249. https://doi.org/10.1504/IJLT.2009.028808

Leiner, D. J. (2024). *SoSci Survey* (Version 3.5.02) [Computer software]. https://www.soscisurvey.de

Machts, N., Chernikova, O., Jansen, T., Weidenbusch, M., Fischer, F., & Möller, J. (2024). Categorization of simulated diagnostic situations and the salience of diagnostic information: Conceptual framework. *Zeitschrift für Pädagogische Psychologie, 38*(1–2), 3–13. https://doi.org/10.1024/1010-0652/a000364

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction, 60*, 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education, 41*(12), 1140–1145. https://doi.org/10.1111/j.1365-2923.2007.02920.x

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Reimer, T., Hoffrage, U., & Katsikopoulos, K. (2007). Entscheidungsheuristiken in Gruppen. [Decision heuristics in groups]. *NeuroPsychoEconomics, 2*(1), 7–29.

Renkl, A. (2014). Lernaufgaben zum Erwerb prinzipienbasierter Fertigkeiten: Lernende nicht nur aktivieren, sondern aufs Wesentliche fokussieren [Learning tasks for acquiring principlebased skills: Not only activating learners, but focusing on the essentials]. In B. Ralle, S. Prediger, M. Hammann, & M. Rothgangel (Eds.), *Lernaufgaben entwickeln, bearbeiten und überprüfen* (pp. 12–22). Waxmann.

Robinson, S. (2023, April 6). *GPTAvatar* [GitHub repository]. GitHub. https://github.com/SethRobinson/GPTAvatar

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction, 83*, 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Schmid, M., & Petko, D. (2020). Technological Pedagogical Content Knowledge als Leitmodell medienpädagogischer Kompetenz. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung, 17* (Jahrbuch Medienpädagogik), 121–140. https://doi.org/10.21240/mpaed/jb17/2020.04.28.X

Seidel, T., Stürmer, K., Schäfer, S., & Jahn, G. (2015). How preservice teachers perform in teaching events regarding generic teaching and learning components. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(2), 84–96. https://doi.org/10.1026/0049-8637/a000125

Shumanov, M., & Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in Human Behavior, 117,* 106627. https://doi.org/10.1016/j.chb.2020.106627

Siegle, R. F., Schroeder, N. L., Lane, H. C., & Craig, S. D. (2023). Twenty-five years of learning with pedagogical agents: History, barriers, and opportunities. *TechTrends, 67*(5), 851–864. https://doi.org/10.1007/s11528-023-00869-3

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in medicine: Knowledge acquisition of conceptual, strategic, and conditional diagnostic knowledge. *Learning and Instruction, 21*(6), 824–834. https://doi.org/10.1016/j.learninstruc.2011.03.004

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Tönnissen, L., & Hövel, D. C. (2022). Die Analyse mit der ICF-CY: Hilfsmittel für einen multifaktoriellen Zugang zur Planung des sozial-emotionalen Lernens? [Analysis using the ICF-CY: Tool for a multifactorial approach to planning social-emotional learning?]. *Emotionale und soziale Entwicklung in der Pädagogik der Erziehungshilfe und bei Verhaltensstörungen: ESE, 4*(4), 132–142. https://doi.org/10.25656/01:24720

van Ophuysen, S., & Behrmann, L. (2015). Die Qualität pädagogischer Diagnostik im Lehrerberuf: Anmerkungen zum Themenheft „Diagnostische Kompetenzen von Lehrkräften und ihre Handlungsrelevanz" [The quality of pedagogical diagnostics in the teaching profession: Annotations to the special issue "Teachers' diagnostic competences and their practical relevance"]. *Journal for Educational Research Online, 7*(2), 82–98. https://doi.org/10.25656/01:11491

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence, 4,* 654924. https://doi.org/10.3389/frai.2021.654924

Zellner, J., Ebenbeck, N., & Gebhardt, M. (2024). Entwicklung digitaler Simulationsspiele mit integrierten Entscheidungsbäumen zur Förderung der diagnostischen Entscheidungskompetenzen in der sonderpädagogischen Lehrkräfteausbildung [Development of digital simulation games with integrated decision trees to promote diagnostic decision-making competence in special education teacher training]. *Qfl – Qualifizierung für Inklusion, 6*(2). https://doi.org/10.21248/Qfl.162

Zhao, Z., Yin, Z., Sun, J., & Hui, P. (2024). Embodied AI-guided interactive digital teachers for education. In *SIGGRAPH Asia 2024 Educator's Forum (SA '24)* (Article 6, pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3680533.3697070