*Journal of Computer Assisted Learning*

WILEY

Journal of Computer Assisted Learning

ORIGINAL ARTICLE OPEN ACCESS

# AI-Based Adaptive Feedback in Simulations for Teacher Education: An Experimental Replication in the Field

Elisabeth Bauer[1,2] | Michael Sailer[1] | Frank Niklas[3] | Samuel Greiff[4] | Sven Sarbu-Rothsching[5] | Jan M. Zottmann[5] | Jan Kiesewetter[5] | Matthias Stadler[5] | Martin R. Fischer[5] | Tina Seidel[2] | Detlef Urhahne[6] | Maximilian Sailer[6] | Frank Fischer[3]

[1]Learning Analytics and Educational Data Mining, University of Augsburg, Augsburg, Germany | [2]School of Social Sciences and Technology, Technical University of Munich, Munich, Germany | [3]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany | [4]Centre for International Student Assessment, Technical University of Munich, Munich, Germany | [5]Institute of Medical Education, LMU University Hospital, LMU Munich, Munich, Germany | [6]Faculty of Social and Educational Sciences, University of Passau, Passau, Germany

**Correspondence:** Elisabeth Bauer (elisabeth.bauer@uni-a.de)

**Received:** 25 June 2023 | **Revised:** 22 December 2024 | **Accepted:** 6 January 2025

**Funding:** This work was supported by the Deutsche Forschungsgemeinschaft, DFG FOR 2385 and Bundesministerium für Bildung und Forschung, 16DHL1040.

**Keywords:** adaptivity | artificial intelligence | feedback | learning analytics | natural language processing | personalisation | simulation-based learning

## ABSTRACT

**Background:** Artificial intelligence, particularly natural language processing (NLP), enables automating the formative assessment of written task solutions to provide adaptive feedback automatically. A laboratory study found that, compared with static feedback (an expert solution), adaptive feedback automated through artificial neural networks enhanced preservice teachers' diagnostic reasoning in a digital case-based simulation. However, the effectiveness of the simulation with the different feedback types and the generalizability to field settings remained unclear.

**Objectives:** We tested the generalizability of the previous findings and the effectiveness of a single simulation session with either feedback type in an experimental field study.

**Methods:** In regular online courses, 332 preservice teachers at five German universities participated in one of three randomly assigned groups: (1) a simulation group with NLP-based adaptive feedback, (2) a simulation group with static feedback and (3) a no-simulation control group. We analysed the effect of the simulation with the two feedback types on participants' judgement accuracy and justification quality.

**Results and Conclusions:** Compared with static feedback, adaptive feedback significantly enhanced justification quality but not judgement accuracy. Only the simulation with adaptive feedback significantly benefited learners' justification quality over the no-simulation control group, while no significant differences in judgement accuracy were found.

Our field experiment replicated the findings of the laboratory study. Only a simulation session with adaptive feedback, unlike static feedback, seems to enhance learners' justification quality but not judgement accuracy. Under field conditions, learners require adaptive support in simulations and can benefit from NLP-based adaptive feedback using artificial neural networks.

## 1 | Introduction

Advances in artificial intelligence (AI)—especially in the context of natural language processing (NLP), as illustrated by large language models and artificial neural networks for automatic text analysis—offer new capacities and opportunities for education (Dai and Ke 2022; Kasneci et al. 2023; Ninaus and Sailer 2022). Among these opportunities is the automation

**Summary**

- What is currently known about this topic?
  - In laboratory settings, simulation-based learning with adaptive feedback can yield significant benefits for learning complex reasoning skills, such as preservice teachers' diagnostic reasoning, compared with simulation-based learning with static feedback.
  - Compared with static feedback, adaptive feedback enhances diagnostic justification skills but not diagnostic judgement skills in a digital simulation.
  - As a basis for adaptive feedback, NLP, using artificial intelligence, facilitates the automation of real-time learning analytics for the formative assessment of learners' written reasoning.
- What does this paper add?
  - Under the field conditions of online courses, we replicated the finding that simulation-based learning with automatic adaptive feedback, compared with static feedback, facilitates preservice teachers' diagnostic justification quality but not their diagnostic judgement accuracy.
  - Compared with a no-simulation control group, simulation-based learning showed positive effects on learners' justification quality only in combination with adaptive feedback and not with static feedback.
  - The effectiveness of single simulation sessions with adaptive feedback may depend on the type of targeted learning outcome, with effects potentially increasing through repeated simulation-based learning.
- Implications for practice/or policy
  - When learning with complex simulated cases, adaptive support seems to play a crucial role in learners achieving relevant learning goals.
  - In higher education courses, artificial intelligence, particularly NLP, enables the implementation of real-time learning analytics to automate formative assessments and adaptive feedback.

of the formative assessment of written responses as a basis for providing adaptive feedback, for example, in simulation-based learning to facilitate diagnostic skill development in teacher education (Sailer et al. 2023). As an essential prerequisite for providing students with effective instruction, adequate learning support and equal opportunities, diagnostic reasoning is considered a core practice of teaching (Grossman 2021). Simulation-based learning has proven to offer opportunities for acquiring skills (Chernikova et al. 2020) and approximating practice situations in teacher higher education programmes (Grossman et al. 2009); however, preservice teachers might need elaborated feedback adapted to their individual learner characteristics (e.g., their current state of knowledge and skills, as indicated by their reasoning performance) to benefit optimally from simulation-based learning.

In a recent laboratory study, Sailer et al. (2023) found that, compared with a simulation with static feedback (i.e., an expert solution), a simulation with adaptive feedback—automated through artificial neural networks—had some positive effects on preservice teachers' diagnostic reasoning. However, the replicability of these effects in uncontrolled field settings and the effectiveness of simulation-based learning with either type of feedback compared with no simulation-based learning remained unclear. Therefore, we conducted a field experiment in online teacher education courses at five German universities to assess (a) whether the findings of Sailer et al. (2023) were replicable under field conditions and (b) whether a single session of simulation-based learning with static feedback or with adaptive feedback had positive effects on learners' reasoning performance, compared with no simulation-based learning.

## 1.1 | Teachers' Diagnostic Reasoning

Diagnostic reasoning about students' thinking and development is considered a core professional practice in teaching and teacher education (Grossman 2021). Teachers' diagnostic reasoning is broadly defined as the goal-oriented collection and integration of information to reduce uncertainty when making educational decisions (Heitzmann et al. 2019). As such, diagnostic reasoning overlaps considerably with concepts such as professional vision, which focuses on teachers' perception and interpretation processes in the classroom (Kramer et al. 2021; Lachner, Jarodzka, and Nückles 2016), and teachers' assessment, which aims to determine a student's current learning state and further directions for learning and development (Black and Wiliam 2009; Herppich et al. 2018). Teachers require diagnostic reasoning in various contexts, such as identifying students' subject-specific understanding (Kron et al. 2021), interpreting students' learning processes in different classroom situations (Kramer et al. 2021) and identifying students with varying motivational–affective and cognitive learning prerequisites (Kosel, Wolter, and Seidel 2021). Among the learning prerequisites most crucial for providing equal opportunities are students' specific learning difficulties, such as a reading disorder, which must be identified at an early stage to avoid disadvantaging affected students throughout their educational careers (Colenbrander, Ricketts, and Breadmore 2018).

Teachers' diagnostic reasoning is conceptualised and researched primarily with regard to diagnostic judgement accuracy—the adequacy of the assessment of student characteristics, such as learning prerequisites, processes and outcomes—because diagnostic judgement accuracy is key to providing students with effective instruction, adequate learning support and equal learning opportunities (Artelt and Rausch 2014; Machts et al. 2016; Urhahne and Wijnia 2021). For example, a primary teacher may be confronted with a student that has significant reading problems, requiring a judgement about the severity and causes of the problems, such as whether the student simply lacks exercise or whether the problems may be due to clinical reasons, such as a reading disorder. In many countries, finalising a diagnostic judgement in the area of reading or spelling disorders is the responsibility of a psychologist or a specialised educational psychologist rather than a teacher (e.g., Colenbrander, Ricketts, and Breadmore 2018); however, teachers are on the front line and therefore play a key role in identifying the early symptoms of a students' specific learning difficulties. Teachers sometimes hold misconceptions and

feel insecure about their knowledge of reading and spelling disorders, which is predicted by the amount of relevant prior education (Peltier et al. 2022). Research on diagnostic reasoning in teacher education and other fields suggests that diagnostic judgement accuracy relies on case-specific diagnostic knowledge stored in well-organised and accessible knowledge structures that are flexible and applicable to reasoning about familiar and unfamiliar cases (Boshuizen, Gruber, and Strasser 2020; Kolovou et al. 2021; Norman et al. 2017). This knowledge includes possible diagnostic activities (e.g., evaluating evidence and drawing conclusions) and evidence that indicates and differentiates between possible diagnoses (Bauer, Sailer, Kiesewetter, Fischer, and Fischer 2022).

Especially in circumstances when teachers need to collaborate, for example, with a specialised educational psychologist, teachers' diagnostic justification quality—that is, whether they can explain an initial diagnostic judgement by providing relevant supportive evidence—is another important dimension of their diagnostic reasoning (Sailer et al. 2023). For example, if a teacher assumes that a student may be affected by a reading disorder, the teacher needs to justify their judgement by explaining relevant evidence related to symptoms, such as the student's low reading accuracy, poor reading comprehension and possibly severe spelling difficulties. In addition to the case-specific diagnostic knowledge required when reasoning about a student case, diagnostic justification quality may be enhanced by knowledge of how to adequately organise and present the reasoning. The role of this second type of knowledge is supported by prior research indicating that preservice teachers who make an accurate diagnostic judgement do not necessarily perform well when asked to justify this judgement with relevant supportive evidence (Bauer, Sailer, Kiesewetter, Fischer, and Fischer 2022). Therefore, to facilitate the development of preservice teachers' diagnostic reasoning skills regarding students' specific learning difficulties, we consider diagnostic judgement accuracy and justification quality as two critical dimensions of diagnostic reasoning development in teacher education.

### 1.2 | Simulation-Based Learning With Feedback for Facilitating Diagnostic Reasoning

Approximating practice through relevant practice representation is a key method for enhancing reasoning skills, such as diagnostic reasoning, in the context of teacher higher education programmes (Grossman et al. 2009; Helleve, Eide, and Ulvik 2023). Processing a variety of practice representations can help preservice teachers structure their knowledge in an accessible way, thereby increasing knowledge applicability and reasoning flexibility (Boshuizen, Gruber, and Strasser 2020; Jossberger, Breckwoldt, and Gruber 2022). Empirical studies and a meta-analysis suggest that simulation-based learning can be an effective learning approach for achieving these ends (Chernikova et al. 2020).

Simulations are simplified yet valid representations of professional practice situations that include features that learners can manipulate (F. Fischer et al. 2022; Sauvé et al. 2007). Simulations provide opportunities for authentic learning in higher education (Codreanu et al. 2020; Jossberger, Breckwoldt, and Gruber 2022; Nachtigall, Shaffer, and Rummel 2022), yet avoid the potential risks and costs of placing novice learners in actual practice settings (De et al. 2019; Kaufman and Ireland 2016). In a simulation environment, preservice teachers may experience similar reasoning difficulties as in real professional situations; however, within simulations, preservice teachers' impasses and mistakes become opportunities for reflection and learning instead of causing harm to students (Dieker et al. 2014; Heitzmann et al. 2023). In higher education, digital simulations have gained particular attention (Bradley and Kendall 2014; Gegenfurtner, Quesada-Pallarès, and Knogler 2014; Thompson et al. 2019) because they facilitate the standardisation of individual simulated situations, offer to present a variety of relevant simulated situations and provide opportunities to practice repeatedly with large numbers of learners (Kaufman and Ireland 2016). However, despite various options for adjusting the demands of the learning task in simulations (e.g., focusing on a few core activities), simulation-based learning can still be challenging for learners because they must apply their knowledge and skills to complex professional situations and tasks (F. Fischer et al. 2022; Frerejean et al. 2023; Machts et al. 2024). By doing so, learner characteristics, such as learners' prior task-relevant knowledge and skills (Hetmanek et al. 2018; Kolovou et al. 2021), self-regulation (van der Graaf et al. 2022) and task-related interest and motivation (Kron et al. 2022; Nickl et al. 2022), influence learners' task perceptions, task performance and learning outcomes in simulations (Heitzmann et al. 2019).

Due to the complexity of simulations, instructional support, such as feedback, is considered an integral part of effective simulation design (Issenberg et al. 2005). Feedback can be defined as information communicated to a learner that is intended to modify his or her thinking or behaviour to improve learning (Shute 2008). Meta-analytic results have supported the role of feedback in the effectiveness of simulation-based learning (Hatala et al. 2014). Especially learners with low prior knowledge do not seem to benefit from simulation-based learning without instructional support; the lack of feedback information paired with a lack of sufficient knowledge and skills to self-assess their own task processing can result in cognitive exhaustion without achieving significant learning gains (Kiesewetter et al. 2020). However, although feedback is generally considered effective for supporting learning, the effectiveness can vary based on the detail of feedback information, the degree of learner guidance and the demands of the learning task. Elaborated feedback types that present high degrees of information for supporting task processing are considered to be most effective for many learning tasks (Hattie and Timperley 2007; Narciss et al. 2014). The positive effect of such elaborated feedback, compared with less elaborated or no feedback, was supported by several meta-analyses, especially with regard to cognitive learning outcomes and demanding learning tasks (Hattie and Timperley 2007; Wisniewski, Zierer, and Hattie 2020). Therefore, we assume that elaborated feedback is needed to support the learning of diagnostic reasoning (i.e., a cognitive learning outcome) using simulations (i.e., a demanding learning task). However, despite the evidence suggesting the need for elaborated feedback in simulation-based learning, further research on optimal feedback design may help to exploit its effectiveness and to understand how much guidance feedback needs to provide to ensure effective learning with simulations.

### 1.3 | Feedback Adaptivity and Formative Assessment in Simulation-Based Learning

Feedback in digital learning environments often varies in its degree of adaptivity and personalisation, that is, whether and how feedback information is tailored to guide learners according to their needs, based on automatically recorded learner data (Bernacki, Greene, and Lobczowski 2021; Kucirkova, Gerard, and Linn 2021; Plass and Pawar 2020; Tetzlaff, Schmiedek, and Brod 2021; Van et al. 2023).

Specifically, technological advancements, especially in the areas of AI and automatic data analysis, have increased interest in adaptive feedback at a micro-level. Micro-adaptive feedback involves adjustments based on a fine-grained automatic analysis of a learner's performance shown in individual learning tasks (Plass and Pawar 2020; Tetzlaff, Schmiedek, and Brod 2021). Supporting learners micro-adaptively during simulation-based learning therefore requires a detailed formative assessment that compares their current performance to a desired learning or performance goal and identifies options for reducing discrepancies between current and goal states. Accordingly, feedback derived from formative assessment involves a comparison of current performance (feed back) with the target performance or learning goal (feed up), and suggests further steps for improvement (feed forward; Hattie and Timperley 2007). For example, micro-adaptive feedback can address preservice teachers' reasoning difficulties in simulated cases by targeting specific knowledge gaps, such as identifying the most relevant evidence to indicate and differentiate possible diagnoses in a particular case (Sailer et al. 2023).

Nevertheless, digital learning environments designed to enhance complex reasoning often rely on static feedback, such as expert solutions, rather than adaptive feedback. While expert static feedback provides elaborated information about optimal task processing (Narciss et al. 2014; Shute 2008), it poses learners the challenge of comparing current and target performance and identifying options for improvement, which requires them to engage in formative self-assessment (Black and Wiliam 2009). Such self-assessment demands a learner's substantial active engagement and self-regulation, the effectiveness of which varies depending on the difficulty of the learning task and the learners' cognitive, metacognitive and motivational–affective prerequisites (Black and Wiliam 2009; Hattie and Timperley 2007; Ifenthaler, Schumacher, and Kuzilek 2023; Nicol 2021). For learners with favourable prerequisites (relative to task difficulty), formative self-assessment—prompted, for example, by expert static feedback—might promote additional reflection and learning; however, learners with less advantageous learning prerequisites might struggle to engage in effective self-assessment (Hattie and Timperley 2007; Narciss et al. 2022; Nicol 2021; Panadero, Brown, and Strijbos 2016).

Compared with static feedback, adaptive feedback, grounded in formative assessment, personalises information to increase feedback specificity and accessibility (Bimba et al. 2017; Shute 2008). Adaptivity may amplify cognitive, metacognitive and motivational–affective benefits of elaborated feedback, such as enhancing knowledge and skill acquisition (Mertens, Finn, and Lindner 2022; Wisniewski, Zierer, and Hattie 2020), reducing the risk of cognitive resource depletion (Chen et al. 2018; Fyfe, DeCaro, and Rittle-Johnson 2015), facilitating self-regulation and metacognitive strategy use (Kuklick, Greiff, and Lindner 2023; Lim et al. 2021) and having favourable effects on motivation and affect (Wisniewski, Zierer, and Hattie 2020; L. Zheng, Zhong, and Niu 2021), at least if learners sufficiently understand and agree to the feedback (Harks et al. 2014; Nicol 2021). By accommodating learners with different learning prerequisites (i.e., high and low manifestations of different learning prerequisites), these advantages likely make adaptive feedback favourable for a broad range of learners. Studies have supported the idea that automatic adaptive feedback has positive effects, for example, on the essay writing performance of secondary school students in a pre–post comparison (Butterfuss et al. 2022); on secondary-school students' scientific argumentation writing, especially when task-specific information is included (Zhu, Liu, and Lee 2020); on exam performance and self-regulated learning behaviour in undergraduate biological sciences learning compared with no feedback (Lim et al. 2021); on perceived helpfulness and reflection during essay writing in a psychology undergraduate course when combined with tutor feedback compared with tutor feedback alone (Gombert et al. 2024); and on outcomes, behaviours and emotions during collaborative learning (L. Zheng, Zhong, and Niu 2021). In their study of undergraduate programming learning, Ahmed et al. (2020) found that automatic adaptive feedback, compared with tutor feedback, increased learners' efficiency in resolving their errors, but the performance benefit seen in the learning phase disappeared in the delayed post-test where no feedback was provided. However, it is crucial that the benefits of adaptive feedback extend beyond a learning phase by not only optimising behaviour during learning but also maintaining performance benefits in unsupported settings, ranging from post-test scenarios to approximating real-life practice (Butler, Godbole, and Marsh 2013; Grossman et al. 2009).

### 1.4 | NLP for Automating Adaptive Feedback on Diagnostic Reasoning

Automating adaptive feedback for learning and reasoning processes can require sophisticated real-time learning analytics methods, such as NLP (Cavalcanti et al. 2021; Dai and Ke 2022). In particular, NLP which employs deep learning techniques has the potential to automate feedback on reasoning tasks at the level of micro-adaptivity (Kasneci et al. 2023; Ninaus and Sailer 2022).

Recently, transformer-based models—a specific type of artificial neural network, such as bidirectional encoder representations from transformers (BERT) and generative pretrained transformers (GPT)—introduced a new era of NLP capacities for extracting linguistic patterns and contextual nuances, especially through extensive pretraining on large corpora (i.e., large language models). These models stand out for their ability to parse complex textual inputs, such as learners' reasoning; however, to ensure feedback quality, these capacities must be thoroughly researched and harnessed before being used for educational feedback (Jansen et al. 2024). This requirement is particularly

pronounced in higher education, where learning is linked intrinsically to specialised knowledge in specific professional fields and situations.

Examples of well-established and tested algorithms for analysing written reasoning data are conditional random fields (CRFs) and artificial neural networks, specifically recurrent neural networks (RNNs). CRFs have been used to classify segments of preservice teachers' reasoning into diagnostic activities, but they rely heavily on input feature quality (Daxenberger et al. 2018). RNNs, including long short-term memory (LSTM) networks and bidirectional RNNs (Bi-RNNs), can process the sequential nature of reasoning and argumentation. LSTMs efficiently manage long-term dependencies in one direction, whereas Bi-RNNs provide context in both sequence directions. For example, Habernal and Gurevych (2016) used a combination of LSTMs and Bi-RNNs for argument mining in text data from online debate portals. Such ensemble models combine the strengths of each component to achieve superior performance. Similarly, Schulz, Meyer, and Gurevych (2019) combined a bidirectional LSTM network with a CRF output layer (BiLSTM-CRF) to automatically segment and classify diagnostic activities in preservice teachers' written diagnostic reasoning. Compared with Daxenberger et al.'s (2018) CRF, this ensemble model significantly improved analytic performance and offers a robust framework for automatically analysing preservice teachers' written reasoning in real-time. Pretrained models can then automate standardised, high-quality feedback by analysing learners' reasoning input and adaptively combining predefined feedback paragraphs (Pfeiffer et al. 2019).

A recent laboratory study tested the ability of a feedback system employing the BiLSTM-CRF introduced by Schulz, Meyer, and Gurevych (2019) to provide automatic adaptive feedback on written task solutions during simulation-based learning to facilitate preservice teachers' diagnostic reasoning about students' specific learning difficulties and behavioural problems (Sailer et al. 2023). Regarding the accuracy of preservice teachers' diagnostic judgements, the results indicated that compared with a simulation with static feedback (i.e., an expert solution), a simulation with adaptive feedback facilitated the judgement accuracy of collaborative dyads of learners. Collaborative learners performed worse in the static feedback condition while collaborative learners who received adaptive feedback in the simulation did not differ from individual learners, who achieved similar judgement accuracy in the adaptive or static feedback condition during the learning phase and in a post-test. Based on the findings, it remains unclear whether the simulation using either feedback type effectively improved preservice teachers' judgement accuracy or whether neither intervention was able to enhance judgement accuracy within a single simulation session. Since diagnostic judgement is a compiled outcome of fine-grained reasoning processes, the simulation might have positive effects in combination with both feedback types; however, a single simulation session that includes only a few cases may not sufficiently broaden learners' case-specific diagnostic knowledge to enhance judgement accuracy effectively across different cases (Boshuizen, Gruber, and Strasser 2020; Kolovou et al. 2021; Norman et al. 2017). Consequently, the laboratory study does not identify whether a single simulation session can significantly benefit preservice teachers' diagnostic judgement accuracy. Further research is needed to determine the effectiveness of a brief simulation-intervention with either type of feedback compared with no simulation-based learning.

Nevertheless, the laboratory study, in contrast to the findings on judgement accuracy, found that adaptive feedback had a large positive effect on the quality of preservice teachers' diagnostic justification during both the learning phase and the post-test. Similarly, the initial results of a parallel field study in medical education indicated that in a post-test, automatic adaptive feedback in simulations had positive effects on medical students' justification quality (Sarbu-Rothsching et al. 2022). Unlike the compiled reasoning outcome of judgement accuracy, learners' justifications may more closely reflect initial improvements in learners' reasoning skills as they give more insights into the learners' reasoning processes. Additionally, knowledge of how to adequately organise and present the reasoning—suggested as a factor that facilitates justification quality—may be less specific to each case and more transferable across different simulated cases (Bauer, Sailer, Kiesewetter, Fischer, and Fischer, 2022; Hetmanek et al. 2018). Nevertheless, it is uncertain whether or not static feedback provided sufficient guidance to enhance preservice teachers' diagnostic justifications in simulation-based learning. To assess the effectiveness of simulation-based learning with static feedback for enhancing preservice teachers' diagnostic justifications, a comparison with no simulation-based learning is needed.

Collectively, these findings indicate that simulations with NLP-based adaptive feedback, which communicates an automated formative assessment, can enhance complex reasoning outcomes, such as the quality of preservice teachers' diagnostic justifications, at least in comparison to a simulation with static feedback and in a well-controlled laboratory setting. It remains unclear to what extent this effect is transferable to an uncontrolled field setting, such as online courses in teacher education programmes, during which students are unobserved and might process the learning tasks differently than they would in a controlled laboratory setting. Whether a single simulation session with either type of feedback can significantly benefit preservice teachers' diagnostic judgement accuracy and whether a single simulation session with static feedback can enhance preservice teachers' diagnostic justification quality compared with no simulation-based learning also remains unknown.

### 1.5 | The Present Field Study

Considering the need for field studies and replication research on the effects of educational interventions (e.g., Makel and Plucker 2014), the first aim of the present study was to conduct an experimental replication study that investigated whether the pattern of results indicating the superiority of automatic adaptive over static feedback in simulation-based learning found in the previous laboratory study (Sailer et al. 2023) would replicate under field conditions. Specifically, the present study focuses on online courses in teacher education programmes at five German universities to investigate the effects of simulation-based learning combined with automatic adaptive or static feedback in an individual learning setting. To determine whether the assumed benefits of adaptive feedback in the simulation

enhance performance not only during the learning phase when feedback is provided, but also in unsupported settings, we investigated judgement accuracy and justification quality at two measurement points: (a) during the learning phase, when learners processed simulated cases and received feedback, and (b) in an unsupported post-test case, where learners received no feedback.

RQ1. The first research question is whether automatic adaptive feedback is more effective than static feedback in enhancing learners' judgement accuracy in simulation-based learning (RQ1a during the learning phase; RQ1b in the post-test). We assumed that adaptive feedback has the potential to enhance the diagnostic knowledge needed to improve learners' judgement accuracy. However, because the evidence of our prior study does not support this assumption, we did not formulate a hypothesis for the first research question but instead explored it as an open question.

RQ2. The second research question is whether automatic adaptive feedback is more effective than static feedback in enhancing learners' justification quality in simulation-based learning (RQ2a during the learning phase; RQ2b in the post-test). We assumed that, compared with static feedback, adaptive feedback might offer more potential benefits in the cognitive, metacognitive and motivational–affective dimensions, thus accommodating a broader range of learners with varying characteristics; this might be particularly relevant to enhancing complex reasoning outcomes, such as providing a high-quality justification in a simulation. The findings of our previous research supported this assumption (Sailer et al. 2023); thus, we hypothesised that automatic adaptive feedback would be more effective than static feedback in enhancing learners' justification quality in simulation-based learning.

In addition to partially replicating the previous laboratory study under field conditions, the second aim of the present field study was to further explore the question of whether exposure to a single simulation session with either adaptive or static feedback would provide a significant advantage for preservice teachers' diagnostic reasoning over a control group that did not involve in simulation-based learning. The effectiveness of simulations may depend on the degree of guidance provided within the feedback, with static feedback providing less specific guidance than adaptive feedback, but potentially sufficient support for effective simulation-based learning. To justify their use in higher education, the effectiveness of simulation sessions with either type of feedback compared with no simulation-based learning needs to be established, ideally under field conditions to ensure ecological validity.

RQ3. Therefore, the third research question is whether, compared with a no-simulation control group, a simulation session with automatic adaptive feedback or with static feedback had a positive effect on learners' judgement accuracy (RQ3a for simulation with adaptive feedback; RQ3b for simulation with static feedback). Following our previous laboratory study, whether preservice teachers' diagnostic judgement accuracy can be increased significantly in a single simulation session with either type of feedback remains unknown. However, this question is of great practical relevance for higher education settings. In general, simulations are considered well suited to training professional skills, such as diagnostic reasoning, if combined with additional support, such as feedback, which provides further information about optimal task processing. Therefore, we hypothesised that compared with the no-simulation control group, the simulation with either feedback type would show a positive effect on learners' judgement accuracy in the post-test.

RQ4. The fourth research question is whether, compared with the no-simulation control group, a simulation session with automatic adaptive feedback or with static feedback had a positive effect on learners' justification quality (RQ4a for simulation with adaptive feedback; RQ4b for simulation with static feedback). From the previous laboratory study, it remains unclear whether simulation-based learning with static feedback effectively enhances preservice teachers' justification quality or whether static feedback does not provide sufficient guidance for effective simulation-based learning, which is relevant to determine for employing simulations with feedback in higher education. For the same reasons outlined above, we hypothesised that compared with the no-simulation control group, the simulation with either feedback type would have a positive effect on learners' justification quality in the post-test.

## 2 | Materials and Methods

### 2.1 | Sample and Design

The initial sample of our field study included $N = 395$ preservice primary and secondary school teachers. The study was conducted in regular online courses in teacher education programmes at five German universities. We used a three-group, between-subjects experimental design. Participants were randomly assigned to one of three groups: (1) simulation-based learning with automatic adaptive feedback, (2) simulation-based learning with static feedback or (3) a no-simulation control group. We had to exclude $n = 63$ participants because they discontinued data collection and did not provide data for one or more of the measurement points (see Procedure section). The final sample included $N = 332$ preservice teachers, with $n = 118$ receiving adaptive feedback, $n = 112$ receiving static feedback and $n = 102$ in the control group. Among them, 281 (85%) were women, 48 (15%) were men and three participants did not indicate their gender. The average age was $M = 22.66$ years (SD = 3.86, min = 18, max = 45, $n = 1$ missing answer) and the average study semester was $M = 4.46$ (SD = 1.87, min = 1, max = 12, $n = 9$ missing answers).

### 2.2 | Learning Environment, Materials and Tasks

In our study, we used CASUS (https://www.instruct.eu/casus/), a computer-based learning environment designed specifically for case-based learning formats, including simulated cases (M. R. Fischer 1999). Preservice teachers received access to simulated cases describing students with reading and writing difficulties (e.g., due to a reading disorder or impaired vision). The two intervention groups processed three consecutive student cases with feedback as an intervention, followed by an additional post-test case without feedback that was answered by

all three experimental groups. In the learning phase, the participants in the intervention groups received either adaptive or static feedback after submitting their answer for a case. Based on the prior study (Sailer et al. 2023), we estimated an approximate processing time of up to 20 min per simulated case, including familiarisation with the case materials, completing two tasks per case and receiving feedback (except in the post-test, which was therefore estimated to take 15 min). Compared with the previous laboratory study, which involved six simulated learning cases on learning difficulties and behavioural problems, we shortened the learning phase in the current field experiment to three learning cases to account for limited course session durations of around 90 min.

At the beginning of the learning phase, learners received a briefing that introduced them to the simulation environment; in a fiction contract, the learners agreed to take on the teacher role and actively familiarised themselves with the simulated cases (see Fink et al. 2021; Jossberger, Breckwoldt, and Gruber 2022). Before engaging with the simulated case, the study participants watched a 10-min input video about teachers' diagnostic responsibilities and processes, as well as diagnostic reasoning concerning students' specific learning difficulties. The purpose of the input video was to activate existing prior knowledge and to compensate for possible differences in prior knowledge among the participants.

We used a document-based digital simulation (Heitzmann et al. 2019) that offered learners access to various materials and information about simulated students' behaviour and learning achievements. Each case began with an introductory statement highlighting the student's most prominent problem, followed by an exploration of multiple information sources: observations of the student's social behaviour and their learning and working behaviour, examples of written exercises, the one or two most recent annual reports, transcripts of conversations with the student, teacher colleagues and the student's parents (see Figure 1).

The simulated cases were designed to suggest a single diagnosis based on the available information, but somewhat ambiguous overall evidence was included to prompt participants to consider alternative explanations. Accurate diagnoses varied between cases and included a reading and spelling disorder, isolated reading disorder, impaired vision and isolated spelling disorder. For a description of a sample case, see Appendix A; all case materials are publicly available in an online repository at https://osf.io/hn7wm/. More information about the design process is provided by Bauer, Sailer, Kiesewetter, Schulz, et al. (2022) and Fink et al. (2021).

Once the participants engaged with a simulated case, they were free to choose the number and sequence of informational sources they consulted. They were then tasked with providing a diagnostic judgement by selecting one diagnosis from a list of 30 options that included both clinical (e.g., reading and spelling disorder, isolated spelling disorder or attention-deficit disorder) and nonclinical diagnoses (e.g., problematic home environment, insufficient schooling). Participants were also required to provide a written explanation justifying their diagnostic judgement.

## 2.3 | Experimental Groups

### 2.3.1 | Simulation-Based Learning With Adaptive Feedback

In the first intervention group, the participants received automatic adaptive feedback on their explanations submitted for each simulated case. We used the NLP-based feedback system implemented and investigated in our prior laboratory study. The

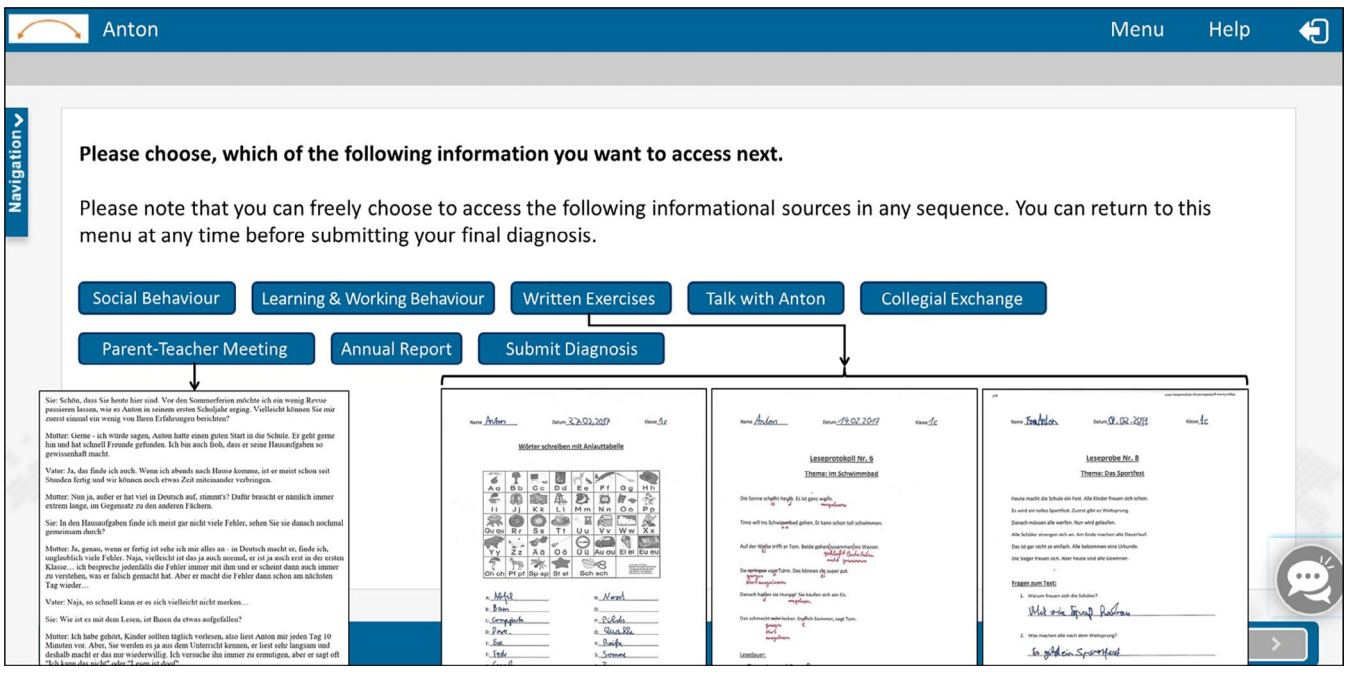


**FIGURE 1** | Examples of simulated student case materials used in the learning phase. Translated from German for publication. See Appendix A for a description of this case.

algorithm used for the feedback system's automatic text analysis was a BiLSTM-CRF, an ensemble model that combines the strengths of Bi-RNNs, LSTMs and CRFs (a more detailed description of the technical feedback implementation is provided by Pfeiffer et al. 2019; details about the algorithm's training and performance have been reported by Pfeiffer et al. 2019 and Schulz, Meyer, and Gurevych 2019). The algorithm's training was based on explanations from $N = 118$ preservice teachers who processed the same cases in an earlier study (Bauer et al. 2020). The training data were labelled manually to identify diagnostic entities (e.g., relevant diagnoses and evidence relevant to their justification, such as reading accuracy, reading comprehension and spelling difficulties; labelling was done by two raters, achieving inter-rater reliabilities ranging from Krippendorff's $\alpha = 0.73$ to $\alpha = 1.00$ across the categories) and diagnostic activities (e.g., generating hypotheses, generating evidence, evaluating evidence and drawing conclusions; labelling was done by four raters, achieving inter-rater reliability of Krippendorff's $\alpha = 0.65$; see Schulz, Meyer, and Gurevych 2019). Accordingly, the algorithm analysed newly written explanations of diagnostic entities and diagnostic activities in real-time.

Based on this analysis, a subset of ~40 predefined and case-specific feedback paragraphs was adaptively activated. One part of the feedback paragraphs addressed the diagnostic activities (i.e., whether the diagnostic activities were identified or missing). The other part of the feedback paragraphs addressed the diagnostic entities (i.e., whether the accurate diagnosis and relevant evidence were correctly explained or missing and, if further diagnoses were mentioned, why these should be rejected). Thereby, the adaptive feedback targeted both the learners' diagnostic judgement and their diagnostic justification in terms of related evidence and diagnostic activities. The adaptively activated feedback paragraphs were sent back to the CASUS learning environment and presented to the learner as an adaptive feedback response (see Figure 2). Analysing learners' submitted explanations and activating the corresponding feedback paragraphs took only a few seconds, so learners received immediate adaptive feedback after submitting an explanation for a learning case. The CASUS interface then showed the adaptive feedback to the learners and additionally allowed learners to click on individual feedback paragraphs that addressed the identified diagnostic activities or entities in the learner's explanation. By clicking on the paragraphs, the identified instances of diagnostic activities or diagnostic entities were highlighted in the learner's submitted explanation to help the learner interpret the feedback (see Figure 2).

### 2.3.2 | Simulation-Based Learning With Static Feedback

Participants in the second intervention group received a case-specific expert solution as static feedback immediately after submitting their diagnostic explanation of a learning case. The static feedback (i.e., expert solutions) and the adaptive feedback were similar in their degree of elaboration and did not differ significantly in terms of content and detail (see Table 1). The static feedback explained relevant diagnostic entities (e.g., relevant diagnoses and related evidence about symptoms, such as reading accuracy, reading comprehension and spelling difficulties) and exemplified diagnostic activities (e.g., generating hypotheses, generating and evaluating evidence and drawing conclusions; see Figure 3). By discussing the accurate diagnosis and related evidence and exemplifying suitable diagnostic activities, the static feedback covered both the diagnostic judgement and diagnostic justification for each case. When receiving static

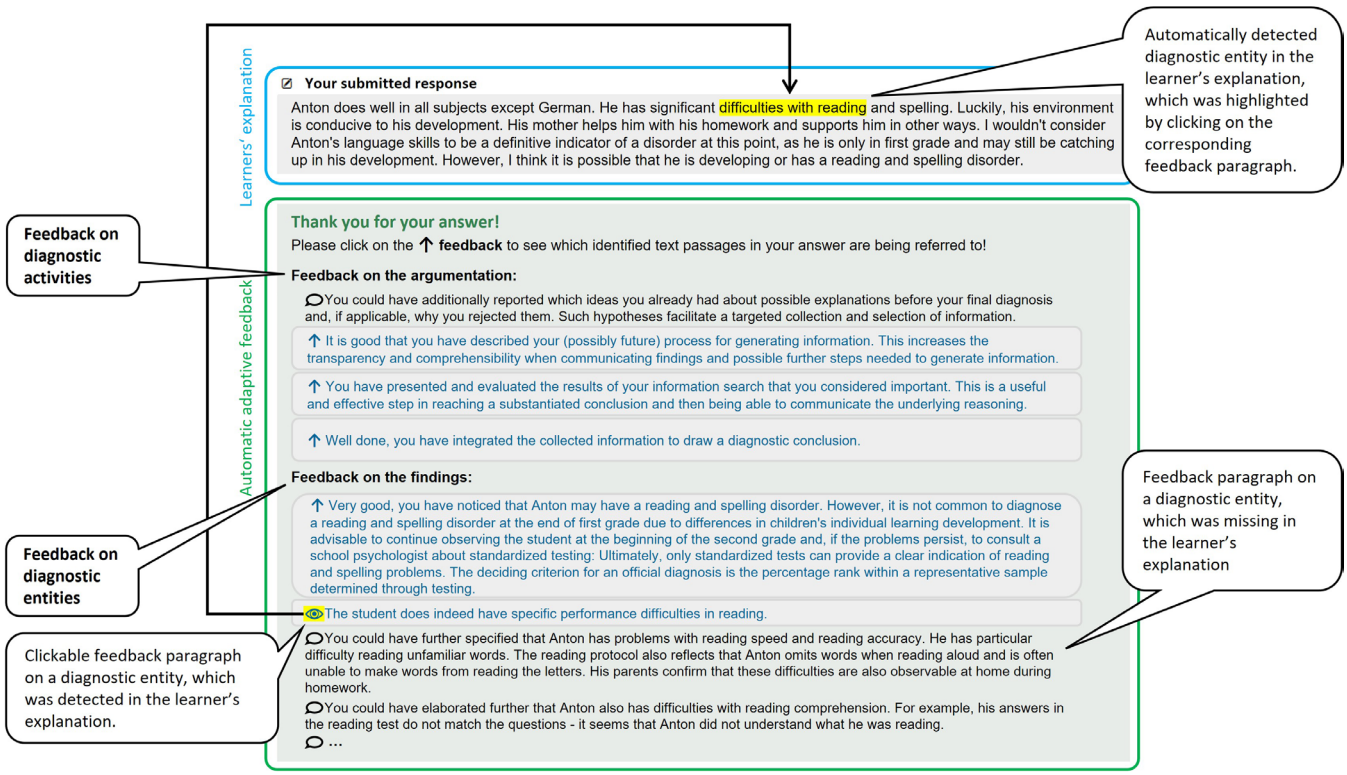


**FIGURE 2** | Automatic adaptive feedback. Translated from German for publication (adapted from Sailer et al. 2023).

**TABLE 1** | Comparison of adaptive and static feedback characteristics in the present study.

| Feedback characteristics | Adaptive feedback | Static feedback |
|---|---|---|
| Feedback timing | Shown immediately after learners' submission of their explanation for a learning case | |
| Feedback content | Includes diagnostic entities and diagnostic activities | |
| Feedback delivery method | Directly addresses whether diagnostic entities were correct, incorrect or missing and whether diagnostic activities were identified or missing | Provides an expert solution that explains relevant diagnostic entities and exemplifies diagnostic activities |

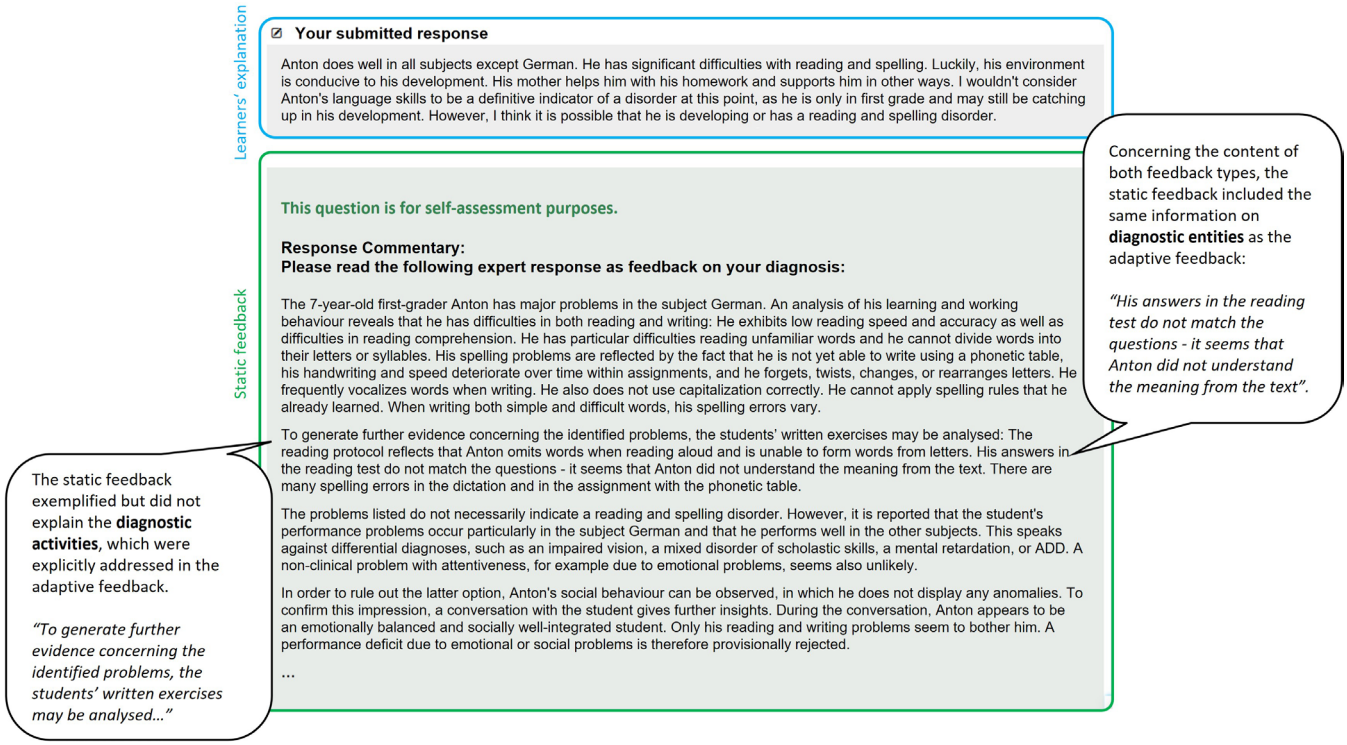


**FIGURE 3** | Static feedback. Translated from German for publication (adapted from Sailer et al. 2023).

feedback, the participants were prompted to compare their own solutions to the expert solutions.

### 2.3.3 | No-Simulation Control Group

Participants in the control group directly processed the post-test, a single simulated student case without feedback. Because this student case was not supported through feedback, we did not expect any significant training effects through this post-test case. As compensation, they subsequently obtained access to the learning phase including the three simulated student cases with adaptive feedback.

## 2.4 | Procedure

At five German universities, we collaborated with teacher educators who agreed to dedicate one of their online course sessions to the simulation tasks and assessments of the present field study. The selected online courses were related to diagnostic reasoning. Before the course sessions, each teacher educator informed the preservice teachers that the session involved processing simulated cases as part of an experimental field study. Participants provided consent for their names and university email addresses to be shared with our research team. Individual user logins for the CASUS learning environment were randomly assigned to the consenting participants, and each login was linked to an experimental variant of the online course. In each course, the user logins were equally distributed across the three experimental groups. One week before the course session, participants received emails that introduced the research team and explained the study's scope and procedure, in addition to providing them with the assigned user logins. A reminder email was sent 1 day before the session.

The participants processed the materials simultaneously, starting at the prescheduled time that their online course session began. The number of participants per session ranged from 10 to 35 preservice teachers, depending on the number of students enrolled in each online course. Participants logged in from various locations and used their own computers without researcher supervision; however, before and during the session, the participants were able to contact the research team via an online chat for assistance with technical issues and questions related to the study procedure.

At the beginning of their course session, the participants logged into the CASUS learning environment, gave informed consent

and completed a demographics questionnaire. Then the intervention groups engaged with the three simulated learning cases and received either adaptive or static feedback (see Learning Environment, Materials and Tasks section). Participants' judgements and justifications in the first case were considered a pre-test measurement because learners gave their answers before receiving any feedback, so the feedback intervention did not influence this measurement. The second and third cases were considered as the learning phase since the participants had already received feedback at this point. After the learning phase, the two intervention groups processed the post-test consisting of one simulated case without receiving feedback. The control group first completed the post-test without feedback and then, as compensation, gained access to the learning phase with adaptive feedback (see Figure 4).

After completing the learning phase and post-test, the participants were informed of the different experimental conditions of the study. Regardless of the initial experimental group to which they were assigned, all participants were offered access to three additional simulated cases (students with behavioural problems) with adaptive feedback. These additional cases were used in a previous laboratory study. The students were given access to the additional cases and were able to process them voluntarily for 4 weeks following the session.

## 2.5 | Data Sources and Measurements

### 2.5.1 | Accuracy of Diagnostic Judgements

We evaluated the participants' diagnostic judgement accuracy using data from the first task in each learning case and the post-test, in which they chose one diagnosis from a list of 30 options. Correct diagnoses were assigned 1 point and incorrect ones received 0 points. We assessed learners' judgement accuracy using three measurements.

Pre-test judgement accuracy measured the participants' judgement accuracy in the first learning case, which was unaffected by the feedback intervention (0 or 1 point).

Judgement accuracy during the learning process was assessed as a mean score of the second and third learning cases (0, 0.5 or 1 point). The two learning process judgements showed a significant correlation (Spearman $r = 0.16$, $p = 0.018$).

Post-test judgement accuracy measured the judgement accuracy of the simulated case in the post-test without subsequent feedback (0 or 1 point).

Data on judgement accuracy in the pre-test and the learning process were available for the two intervention groups ($n = 230$), while the control group processed the learning phase only as compensation following the post-test. Post-test data were collected for all three experimental groups ($n = 332$).

### 2.5.2 | Quality of Diagnostic Justifications

We used the second task in each simulated case during the learning phase and the post-test to assess the participants' justification quality. Using the coding scheme developed for our previous study (see Sailer et al. 2023; the coding scheme for the above example case is shown in Appendix B), we assigned 1 point for each of the six primary supporting pieces of evidence for the respective accurate diagnosis. We considered an explanation containing all six pieces of evidence to be an indicator of high justification quality. Two trained raters independently coded 10% of the post-test explanations with high inter-rater reliability (Cohen's $\kappa = 0.92$). One rater then coded the remaining post-test explanations. The second rater and a third rater independently coded 10% of the written explanations of the learning phase, again achieving high inter-rater reliability (Cohen's $\kappa = 0.91$). The second rater coded the remaining explanations of the learning phase. For further analyses, we differentiated between three measurements of justification quality.

Pre-test justification quality was based on each participant's first learning case before receiving feedback. For each participant, we calculated a mean score of the points achieved for pre-test justification quality; thus, the score ranged from 0 to 1 point. The measurement's internal consistency was relatively low (McDonald's $\omega = 0.50$).

Justification quality during the learning process was measured by calculating the mean of the points achieved for the second

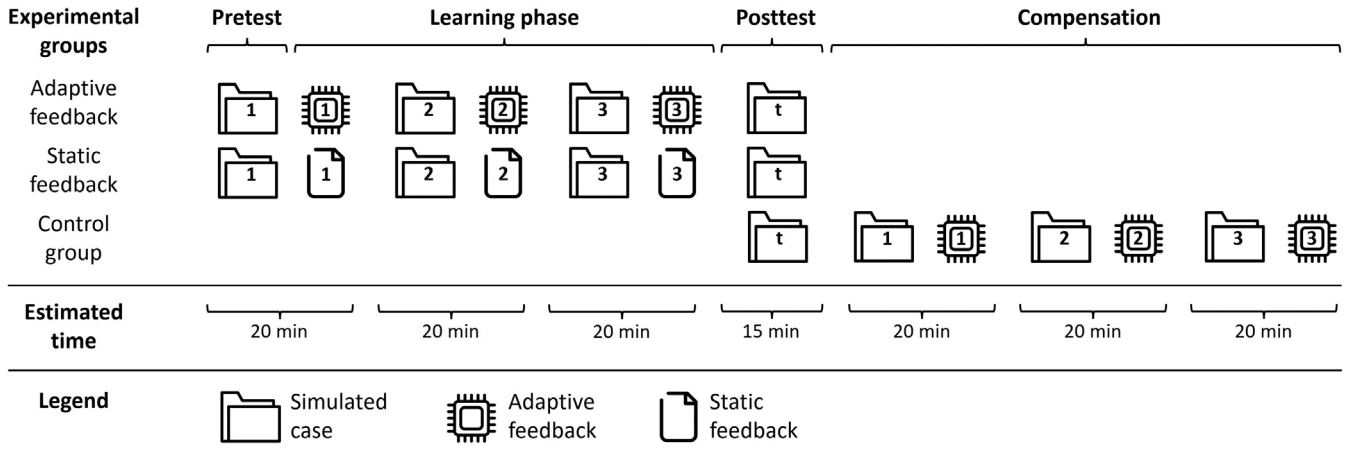


**FIGURE 4** | Study procedure.

and third simulated cases in the learning phase. The score thus ranged from 0 to 1 point. The measurement's internal consistency was acceptable (McDonald's $\omega = 0.73$).

The post-test justification quality measured participants' performance in the post-test case without feedback. Again, we calculated the mean score of the points achieved for post-test justification quality, resulting in a score ranging from 0 to 1 point. The measurement's internal consistency was acceptable (McDonald's $\omega = 0.61$).

Data regarding pre-test justification quality and justification quality during the learning process were available for the two intervention groups ($n = 230$ preservice teachers), while post-test justification quality was analysed for all three experimental groups ($n = 332$ preservice teachers).

### 2.6 | Statistical Analyses

To investigate RQ1a and RQ1b, we conducted a MANCOVA using feedback type as the independent variable, pre-test judgement accuracy as a covariate, and (1a) judgement accuracy in the learning process and (1b) post-test judgement accuracy as dependent variables. Similarly, for RQ2a and RQ2b, we performed a MANCOVA with feedback type as the independent variable, pre-test justification quality as a covariate, and (2a) justification quality in the learning process and (2b) post-test justification quality as dependent variables. For RQ3a and RQ3b, we used a univariate ANOVA with planned contrasts to compare the control group to simulation-based learning with (3a) adaptive feedback and with (3b) static feedback, using post-test judgement accuracy as the dependent variable. Likewise, for RQ4a and RQ4b, we conducted a univariate ANOVA with planned contrasts to compare the control group to simulation-based learning with (4a) adaptive feedback and with (4b) static feedback, using post-test justification quality as the dependent variable. IBM SPSS Statistics 28 was used to perform all analyses, and the alpha level was set to $\alpha = 0.05$.

## 3 | Results

### 3.1 | Randomisation Check for the Intervention Groups

We compared learners' performance on the first learning case to obtain an indicator of their prior knowledge and skills. The feedback intervention did not influence this measurement because the learners submitted their answers before receiving any feedback. Table 2 presents the descriptive results of the learners' judgement accuracy and justification quality for the first learning case. Judgement accuracy was significantly higher in the static feedback group than in the adaptive feedback group, with a small effect size ($t(204.04) = 2.13$, $p = 0.034$, Cohen's $d = 0.278$). No significant difference in justification quality was found between the two intervention groups ($t(228) = 0.49$, $p = 0.626$, Cohen's $d = 0.064$). These results indicate that learners' prior skills and knowledge were not entirely equal; therefore, we used their performance (judgement accuracy and justification quality) in the first learning case as a covariate for subsequent analyses that compared the effects of adaptive and static feedback on learners' judgement accuracy and justification quality, both during the learning process and in the post-test.

### 3.2 | Effects of Feedback Type in Simulation-Based Learning

#### 3.2.1 | Diagnostic Judgement Accuracy

To explore RQ1, we calculated a MANCOVA with learners' judgement accuracy in the first learning case as a covariate (judgement accuracy in the learning process, $F(1,228) = 2.21$, $p = 0.139$, $\eta_p^2 = 0.010$; judgement accuracy in the post-test, $F(1,228) = 7.95$, $p = 0.005$, $\eta_p^2 = 0.034$). Descriptive statistics for learners' judgement accuracy in the learning process and in the post-test are shown in Table 3. There were no significant differences between the judgement accuracy of learners who received adaptive feedback and learners who received static feedback in

**TABLE 2** | Learners' judgement accuracy and justification quality in the first learning case before the feedback intervention.

| | Adaptive feedback | Static feedback |
|---|---|---|
| **Dependent variable** | ***M* (SD)** | ***M* (SD)** |
| Judgement accuracy in the first learning case | 0.86 (0.34) | 0.95 (0.23) |
| Justification quality in the first learning case | 0.39 (0.24) | 0.40 (0.22) |

**TABLE 3** | Learners' judgement accuracy and justification quality in the learning process and post-test (without controlling for the pre-test).

| | Adaptive feedback | Static feedback | Control group |
|---|---|---|---|
| **Dependent variable** | ***M* (SD)** | ***M* (SD)** | ***M* (SD)** |
| Judgement accuracy in the learning process | 0.65 (0.37) | 0.72 (0.31) | — |
| Judgement accuracy in the post-test | 0.85 (0.36) | 0.91 (0.29) | 0.83 (0.37) |
| Justification quality in the learning process | 0.46 (0.42) | 0.41 (0.24) | — |
| Justification quality in the post-test | 0.55 (0.25) | 0.48 (0.23) | 0.47 (0.27) |

the simulation, either in the learning process ($F(1,228) = 1.42$, $p = 0.234$, $\eta_p^2 = 0.006$) or in the post-test ($F(1,228) = 1.12$, $p = 0.290$, $\eta_p^2 = 0.005$). Thus, the effect of adaptive and static feedback on enhancing learners' judgement accuracy did not differ (RQ1a/1b).

#### 3.2.2 | Diagnostic Justification Quality

To investigate RQ2, we again calculated a MANCOVA with the learners' justification quality in the first learning case as a covariate (justification quality in the learning process, $F(1,227) = 115.15$, $p < 0.001$, $\eta_p^2 = 0.339$; justification quality in the post-test, $F(1,227) = 42.62$, $p < 0.001$, $\eta_p^2 = 0.159$). Descriptive statistics for learners' justification quality in the learning process and in the post-test are shown in Table 3. Compared with learners who received static feedback, the justification quality of learners who received adaptive feedback in the simulation was significantly higher in the learning process ($F(1,227) = 4.01$, $p = 0.047$, $\eta_p^2 = 0.017$) and the post-test ($F(1,227) = 6.78$, $p = 0.010$, $\eta_p^2 = 0.029$), with both showing small effect sizes. These results support our hypothesis that adaptive feedback is more effective than static feedback in enhancing learners' justification quality in simulation-based learning (RQ2a/2b).

### 3.3 | Comparison With the No-Simulation Control Group

#### 3.3.1 | Diagnostic Judgement Accuracy

To investigate RQ3, we calculated an ANOVA with planned contrasts. Descriptive statistics for judgement accuracy in the post-test are shown in Table 3. Overall, learners' judgement accuracy did not differ significantly between the three experimental groups ($F(2,329) = 1.59$, $p = 0.205$, $\eta_p^2 = 0.010$). Planned contrasts showed no significant difference between the no-simulation control group and the group that processed the simulation with static feedback ($F(1,329) = 2.73$, $p = 0.099$, $\eta_p^2 = 0.008$); there was also no significant difference between the no-simulation control group and the group that processed the simulation with adaptive feedback ($F(1,329) = 0.09$, $p = 0.760$, $\eta_p^2 < 0.001$). These results do not support the hypothesis that, compared with the no-simulation control group, a single simulation session with either feedback type has a positive effect on learners' judgement accuracy (RQ3a/3b).

#### 3.3.2 | Diagnostic Justification Quality

We also calculated an ANOVA with planned contrasts to investigate RQ4. Descriptive statistics for justification quality in the post-test are shown in Table 3. Learners' justification quality significantly differed between the three experimental groups ($F(2,327) = 3.37$, $p = 0.036$, $\eta_p^2 = 0.020$). Planned contrasts showed no significant difference between the no-simulation control group and the group that processed the simulation with static feedback ($F(1,327) = 0.04$, $p = 0.838$, $\eta_p^2 < 0.001$); however, there was a significant difference between the no-simulation control group and the group that processed the simulation with adaptive feedback ($F(1,327) = 5.32$, $p = 0.022$, $\eta_p^2 = 0.016$). These results do not support the hypothesis that compared with the no-simulation control group, a simulation session with static feedback has a positive effect on learners' justification quality in the post-test (RQ4b); however, the results do support the hypothesis that compared with the no-simulation control group, a single simulation session with adaptive feedback has a positive effect on learners' justification quality (RQ4a).

## 4 | Discussion

In this study, we investigated the effects of digital case-based simulations with NLP-based adaptive feedback on the diagnostic reasoning of preservice teachers enrolled in online courses at five German universities. During the learning phase and in an unsupported post-test, automatic adaptive and static feedback did not show different effects on preservice teachers' judgement accuracy, but adaptive feedback does appear to be more effective than static feedback regarding the quality of diagnostic justification. These findings replicate previous result patterns, including the patterns observed in a previous laboratory study (Sailer et al. 2023) and in a parallel field study about automatic adaptive feedback in medical education simulations (Sarbu-Rothsching et al. 2022).

In addition to replicating a comparison of the two feedback types used within the simulation, we compared two intervention groups (simulation with adaptive feedback and simulation with static feedback) against a no-simulation (and no feedback) control group to assess whether a single simulation session with either feedback type significantly benefits preservice teachers' diagnostic reasoning. This comparison aimed to investigate whether simulations are effective and suitable for use in higher education field settings regardless of the feedback type or if the simulation's effectiveness depends on the specific type of feedback provided. Contrary to our hypothesis (RQ3a/3b), a comparison of the intervention groups to the no-simulation control group revealed no significant benefits associated with the simulation session—regardless of feedback type—on learners' judgement accuracy in the post-test. Regarding justification quality, the simulation with adaptive feedback demonstrated a significant positive effect compared with the no-simulation control group, while the simulation with static feedback did not. We conclude that adaptive feedback is essential for the effectiveness of simulation sessions in higher education field settings, whereas static feedback seems insufficient for effective simulation-based learning.

### 4.1 | Enhancing Diagnostic Reasoning Through Simulations With Adaptive Feedback

Our findings indicate that simulation-based learning with automatic adaptive feedback enhances learners' justification quality but does not increase their judgement accuracy. The automatic adaptive feedback micro-adaptively adjusted to relevant learner characteristics, especially current diagnostic knowledge and reasoning skills as indicated by their diagnostic reasoning outcomes in each simulated case (Plass and Pawar 2020; Tetzlaff,

Schmiedek, and Brod 2021). The adaptive feedback clearly identified whether relevant diagnostic entities (i.e., the indicated diagnoses and supporting pieces of evidence) were correct, incorrect or missing; it also pointed out whether appropriate diagnostic activities (e.g., generating evidence, evaluating evidence, drawing conclusions) were included in the learners' written responses. In all instances, the adaptively selected feedback paragraphs explained how the targeted diagnostic entities and activities were relevant to the task in general or the simulated case in particular. Moreover, to increase the specificity, transparency and interpretability of the automatic analysis (Chaudhry, Cukurova, and Luckin 2022; Shute 2008), the adaptive feedback highlighted the identified relevant parts of the learners' submitted explanations to facilitate feedback interpretation.

In doing so, the adaptive feedback communicated a formative assessment of the learners' current level of diagnostic reasoning performance in the simulated cases and identified specific options for future improvements (Bimba et al. 2017; Black and Wiliam 2009; Hattie and Timperley 2007). As effects on justification quality persist beyond the learning phase, in which learners received feedback, to an unsupported post-test case, in which no feedback was given, we conclude that the positive effect is not limited to optimised behaviour during learning but may include more sustainable cognitive and metacognitive benefits. Compared with static feedback, adaptive feedback, which provides specific information on learners' individual performance, may be more beneficial for knowledge and skill acquisition (Mertens, Finn, and Lindner 2022; Wisniewski, Zierer, and Hattie 2020), less likely to cognitively overwhelm learners (e.g., by comparing their current with a goal performance; Chen et al. 2018; Fyfe, DeCaro, and Rittle-Johnson 2015; Shute 2008) and favourable in terms of learners' self-regulation and metacognitive strategy use (Kuklick, Greiff, and Lindner 2023; Lim et al. 2021). However, the potential roles that these mechanisms and learner agreement with the received feedback (Harks et al. 2014; Nicol 2021) play in simulation-based learning for professional reasoning tasks require further investigation.

In contrast, learners who received static rather than adaptive feedback in the simulation had to engage in formative self-assessment of their current state of performance (compared with an expert solution) and their options for improvements. This formative self-assessment, stimulated through static feedback, can place higher demands on learners' understanding, cognitive resources, active engagement and self-regulation (Black and Wiliam 2009; Ifenthaler, Schumacher, and Kuzilek 2023; Nicol 2021). While this process may offer benefits through additional reflection for learners with favourable prerequisites relative to the general demands of a learning task, learners with less advantageous learning prerequisites might struggle to engage in effective self-assessment (Hattie and Timperley 2007; Nicol 2021; Panadero, Brown, and Strijbos 2016). Especially in simulation-based learning, the demands associated with formative self-assessment exacerbate the demands of applying knowledge and skills to simulated cases (F. Fischer et al. 2022; Frerejean et al. 2023), which may be a challenging combination for many learners. In this study, learners who received static feedback during simulation-based learning showed no evidence of significant benefits to their justification quality compared with the no-simulation control group, whereas the adaptive feedback group showed evidence of a significantly higher justification quality compared with both other groups. Considering the general demands of simulation-based learning itself, adaptive feedback may be considered favourable for a broader range of learners, as it accommodates learners with more and less favourable learning prerequisites by providing more specific guidance. Thus, adaptive support, which spares learners' cognitive and metacognitive learning resources by providing and highlighting personalised feedback information, seems crucial for achieving relevant learning goals in simulations, such as increasing justification quality in the context of diagnostic reasoning. Accordingly, for higher education field settings in teacher education, a single simulation session that may rather be challenging to students, relative to their prior knowledge, only seems to be recommended with adaptive feedback, not with static feedback, as compared with no simulation-based learning. While the effect on justification quality observed in the present field study was small, the previous laboratory study reported a large effect (Sailer et al. 2023). Given the less controlled nature of the field setting, this discrepancy was expected. Participants completed the study unsupervised and from various locations using their own computers. This uncontrolled and unsupervised environment may have included environmental factors such as distractions. However, these conditions authentically reflect typical online university courses (Hollis and Was 2016). Therefore, compared with laboratory settings, the study's authentic setting enhances the ecological validity and generalizability of our findings. The study suggests that while static feedback provides insufficient guidance for simulation-based learning in higher education field settings, adaptive feedback is crucial for single simulation sessions to effectively enhance reasoning skills in university online courses, at least when the simulation is likely to be challenging relative to learners' prior knowledge.

### 4.2 | Effects of Simulation Sessions on Different Types of Reasoning Outcomes

The effectiveness of single simulation sessions seems to depend not only on combining simulations with adaptive feedback but also on the type of reasoning outcome. We found that the simulation session (with either feedback type) had no effect on preservice teachers' judgement accuracy. This result might be explained by judgement accuracy's high reliance on case-specific diagnostic knowledge; that is, the knowledge that helps to make an accurate judgement for one specific student case may not be directly transferable to other student cases (e.g., Kolovou et al. 2021; Stadler, Sailer, and Fischer 2021). The accurate diagnoses for the simulated cases were not identical, and the case information (e.g., relevant evidence and distractor information) varied considerably. The simulated cases in the learning phase and the post-test might have been so specific that the new insights from one simulated case were not necessarily directly transferable to the other simulated cases (Hetmanek et al. 2018). Additionally, judgement accuracy—conceptualised in the present study as being either an adequate or an inadequate assessment of the students' learning difficulties—represents a consolidation of finer-grained reasoning processes and activities (i.e., evaluating various hypotheses based on distinct evidence and drawing overall conclusions; Heitzmann et al. 2019). While a single simulation session might have some effect on these reasoning processes, this effect may not be sufficient to significantly change compiled

reasoning outcomes, such as preservice teachers' judgement accuracy. Therefore, we conclude that the benefits of single simulation sessions might be limited in terms of enhancing compiled reasoning outcomes that rely highly on specific knowledge.

Compared with judgement accuracy, the single simulation session—at least with adaptive feedback—achieved a significant positive effect on preservice teachers' justification quality compared with the no-simulation control group. Formulating high-quality justifications requires case-specific diagnostic knowledge to identify and select relevant supporting evidence (Bauer, Sailer, Kiesewetter, Fischer, and Fischer, 2022). As high justification quality in this study is defined as explaining the evidence most relevant for supporting the accurate diagnostic judgement, the outcome of justification quality may even give a good insight into fine-grained reasoning processes. Thus, compared with judgement accuracy, justification quality might be an outcome relatively proximal to reasoning processes and, thus, a comparably sensitive indicator of initial changes in learners' reasoning achieved in a single simulation session. Moreover, preservice teachers' justification quality may not rely exclusively on case-specific diagnostic knowledge. Recent research suggests that preservice teachers' diagnostic justification might also rely on knowledge about how to organise and present relevant evidence and arguments (Bauer, Sailer, Kiesewetter, Fischer, and Fischer, 2022). This knowledge is likely less specific to each case and thus easier to transfer across different simulated cases (Hetmanek et al. 2018). Accordingly, single simulation sessions with adaptive feedback show a positive effect on outcomes that are relatively proximal to reasoning processes or that, in addition to case-specific knowledge, rely on knowledge and skills that are transferable across different cases.

Since high justification quality in this study is conceptualised as an outcome proximal to reasoning processes, it may indicate the deliberate processing of relevant case information during learners' reasoning. Especially when practiced repeatedly, deliberate processing of relevant case information has a high potential for increasing learners' case-specific diagnostic knowledge (Andriessen and Baker 2014) and, consequently, their judgement accuracy in sufficiently similar cases. Facilitating preservice teachers' engagement with diverse student cases and expanding their diagnostic knowledge may require repeated practice across multiple simulated cases in several simulation sessions. Repeated learning with authentic cases is considered beneficial for expanding and reorganising knowledge into accessible knowledge structures, thereby extending the applicability and flexibility of knowledge to reasoning about familiar and unfamiliar cases (see Jossberger, Breckwoldt, and Gruber 2022; Norman, Young, and Brooks 2007). Therefore, while a single simulation session has some benefits for certain reasoning outcomes, the full potential of simulation-based learning with adaptive feedback may instead be realised when integrated into opportunities for repeated and deliberate practice over longer periods of study (e.g., over a full semester).

### 4.3 | NLP for Automating Micro-Adaptive Learning Support

Our study suggests that micro-adaptive learning support is crucial for the effectiveness of simulation-based learning in complex reasoning tasks in higher education. AI, particularly NLP, can provide this adaptive support by enabling the implementation of real-time learning analytics and automating the formative assessment of written responses (Dai and Ke 2022; Gardner, O'Leary, and Yuan 2021; Kasneci et al. 2023; Ninaus and Sailer 2022). In our study, we used NLP algorithms that were well-established and researched in the context of diagnostic reasoning in teacher education, specifically BiLSTM-CRF (Schulz, Meyer, and Gurevych 2019), to automatically analyse learners' written diagnostic reasoning and adaptively activate and combine predefined feedback paragraphs into an adaptive feedback response. This approach offered three distinct advantages: first, the algorithms were trained and tested on data pertaining to preservice teachers' diagnostic reasoning about the simulated cases used in this study; second, the feedback paragraphs were formulated specifically for the simulated cases, which ensured the quality of specialised learning domain content; third, prior research conducted in a laboratory setting (Sailer et al. 2023) identified the benefits of this approach before its application in online teacher education courses at multiple universities, thus ensuring a systematic and ethical implementation of AI before its adoption in a higher education field setting.

Further enhancing the micro-adaptivity of automatic adaptive feedback could yield additional benefits. The artificial neural networks used in this study did not yet employ the latest advancements in NLP, which have significantly advanced the ability of artificial neural networks to perform various NLP tasks (Kasneci et al. 2023). Transformer models used to build and train large language models process text more efficiently than RNNs, which are limited by sequential processing. By simultaneously processing sequences and dynamically adjusting attention to capture contextual relationships, transformer models may more efficiently analyse preservice teachers' written reasoning, which may facilitate more precise activation of predefined feedback paragraphs. However, further research is needed to determine whether large language models require additional fine-tuning for specialised domains, such as preservice teachers' diagnostic reasoning, to fully realise their potential to provide micro-adaptive learning support in professional learning scenarios.

Besides text analysis, large language models have also enhanced NLP's capabilities in the realm of generative AI, particularly in text generation (e.g., ChatGPT; OpenAI 2023). Harnessing these capabilities to generate uniquely tailored feedback messages instead of adaptively providing predefined feedback paragraphs, might increase the specificity and effectiveness of automatic adaptive feedback. Effectively using these capabilities to automatically generate high-quality adaptive feedback messages in the context of learning with complex simulated cases remains a challenge. Specifically, employing text generation to produce feedback involves unlocking the learning potential within specialised domains and ensuring the pedagogical quality of formative assessment and high-information feedback.

### 4.4 | Limitations and Future Research

The generalizability of our results may be limited due to the pre-test (i.e., the first learning case) and post-test measurements consisting only of single simulated cases. A low number

of measurements are often associated with mediocre measurement reliability, which may explain the low internal consistency of the pre-test justification quality, limiting the systematicity of the measurement of this covariate. Moreover, a low number of measurements potentially limits the observable variance, which might challenge our interpretation that there were no differences in preservice teachers' judgement accuracy. The approach to operationalising judgement accuracy (i.e., letting learners choose from a list of diagnoses as well as the dichotomous rating of the accuracy) might have further limited the observable variance. However, our interpretation of the comparison of adaptive and static feedback regarding both justification quality and judgement accuracy is supported by the consistent results pattern found in the learning phase and the post-test as well as the consistency with the results pattern found in the previous laboratory study. As the study design only included a no-simulation control group, but not a simulation-without-feedback control group, the design is limited as to whether the simulated cases may have contributed to individual participants' outcomes independently of the feedback. However, the results suggest that even the static feedback provided insufficient guidance for learners to achieve significant effects in a single simulation session compared with no simulation-based learning, supporting the assumption that the simulation session would probably not have been effective without any feedback. Furthermore, the generalizability of our interpretation regarding the nonsignificant difference between the control group and the intervention group receiving static feedback in the simulation might be limited by the lack of a pre-test measure for the no-simulation control group. Thus, we cannot rule out that the task-related prior knowledge and skills of the control group might have differed from those of the two intervention groups. Moreover, for the post-test measurement, because the no-simulation control group did not access the learning cases before the post-test, they were less familiar with the simulation-based learning environment than the two intervention groups, potentially affecting the comparability of the control group to the intervention groups. In addition, we cannot completely rule out the possibility that individual participants in the control group may have had a minor training effect by participating in the post-test, which consisted of processing a case in the simulation environment. This was, however, done to achieve high measurement validity. Generally, the number of measurements in this study was limited by the time required to process individual cases and the time limits imposed by the scheduled course sessions. In future research, the number of measurements might be increased by investigating simulations with feedback over the course of several course sessions. Such an extended study duration would also allow us to further investigate the assumption that repeated simulation-based learning with adaptive feedback has the potential to increase preservice teachers' judgement accuracy.

Considering that the artificial neural networks used for the NLP tasks in our study did not yet incorporate transformer-based models (Kasneci et al. 2023), future research might explore whether these algorithms can improve the performance of automatic text analysis and, consequently, the effectiveness of our automatic adaptive feedback system for learning diagnostic reasoning in a simulation. Additionally, investigating the use of text generation in automatic adaptive feedback for simulation-based learning of reasoning skills in specialised professional domains is an important desideratum for future research.

Our current field experiment and previous laboratory experiment focused on the effectiveness of simulation-based learning with automatic adaptive feedback. Further laboratory research might explore how automatic adaptive feedback influences preservice teachers' learning processes to better understand the cognitive, metacognitive and motivational–affective mechanisms that underlie the benefits offered by automatic adaptive feedback in simulations. For this purpose, multimodal data—including, for example, trace data, eye tracking, other physiological measurements, questionnaire data and think-aloud data—might provide rich information and detailed insights (see Fan et al. 2023; Kalyuga and Plass 2018; Raković et al. 2023). These learning processes may depend on learners' understanding and agreement with feedback content (Harks et al. 2014; Nicol 2021), which might moderate the effects of feedback and therefore deserve specific attention. Additionally, potential interactions with learner characteristics might be of interest; for example, learners with low prior task-relevant knowledge and skills or low self-regulation skills might have a comparatively high need for adaptive support and feedback in digital learning environments. Identifying the most relevant learner characteristics and learning processes when learning with case-based simulations and adaptive feedback, as well as identifying relevant indicators of these characteristics and processes, might inform the future design of automatic adaptive feedback systems in terms of which additional data, other than text data, could be involved in adjusting automatic adaptive feedback.

In addition to computer-based feedback, NLP might have additional applications in other types of automatic adaptive learning support, such as supporting learners when providing peer assessment and feedback in digital learning environments (Bauer et al. 2023). Finally, future research might investigate adapting the demands of the simulated task itself, for example, by adjusting the informational complexity of the simulated cases to accommodate learner characteristics, such as learners' prior knowledge and skills (F. Fischer et al. 2022).

## 5 | Conclusion

In this study, we investigated and compared the effects of digital case-based simulations with NLP-based adaptive feedback and static feedback on preservice teachers' diagnostic reasoning in online higher education courses. Based on the quality of diagnostic justifications, we found that simulation-based learning with automatic adaptive feedback can enhance preservice teachers' diagnostic reasoning compared with simulation-based learning with static feedback; these results replicate the results pattern found in a previous laboratory study (Sailer et al. 2023). Compared with the no-simulation control group, only simulation-based learning with adaptive feedback, not with static feedback, had positive effects on preservice teachers' justification quality. These findings emphasise the degree to which adaptivity is essential to providing effective feedback in simulation-based learning (Bimba et al. 2017; Dai and Ke 2022; Plass and Pawar 2020). The study indicates that, in higher education field settings, static feedback does not seem to offer adequate guidance to learners in single simulation sessions, whereas adaptive feedback provides critical support for single simulation sessions to effectively enhance reasoning skills,

particularly when the simulation may present a challenge relative to learners' prior knowledge. Adaptive feedback on the reasoning processes captured in written task solutions can be automated by NLP, which employs AI combined with deep learning techniques (Cavalcanti et al. 2021). Given the findings of this study, integrating NLP-based adaptive feedback within simulation-based learning shows great potential for enhancing higher education, potentially redefining how educational technology supports the development of reasoning skills in future generations of professionals.

### Author Contributions

**Elisabeth Bauer:** conceptualization, methodology, data curation, investigation, formal analysis, visualization, writing – review and editing, writing – original draft, software. **Michael Sailer:** conceptualization, methodology, investigation, writing – review and editing, resources, visualization, software. **Frank Niklas:** investigation, writing – review and editing. **Samuel Greiff:** writing – review and editing, resources. **Sven Sarbu-Rothsching:** visualization, writing – review and editing, investigation. **Jan M. Zottmann:** writing – review and editing. **Jan Kiesewetter:** writing – review and editing. **Matthias Stadler:** writing – review and editing. **Martin R. Fischer:** writing – review and editing, conceptualization, funding acquisition. **Tina Seidel:** writing – review and editing, resources. **Detlef Urhahne:** investigation, writing – review and editing. **Maximilian Sailer:** investigation, writing – review and editing. **Frank Fischer:** conceptualization, methodology, writing – review and editing, resources, funding acquisition, supervision, investigation.

### Acknowledgements

This research was supported by the German Federal Ministry of Research and Education (FAMULUS-Project 16DHL1040) and the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG FOR 2385).

During the writing and revision of this article, the authors used ChatGPT 4 to revise the draft for readability and language. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

### Ethics Statement

The study was approved by the Medical Faculty's Ethics Committee of Ludwig-Maximilians-Universität München (no.17–249).

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### References

Ahmed, U. Z., N. Srivastava, R. Sindhgatta, and A. Karkare. 2020. "Characterizing the Pedagogical Benefits of Adaptive Feedback for Compilation Errors by Novice Programmers." In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Software Engineering Education and Training*, edited by G. Rothermel and D.-H. Bae, 139–150. New York, NY, USA: ACM. https://doi.org/10.1145/3377814.3381703.

Andriessen, J., and M. Baker. 2014. "Arguing to Learn." In *Cambridge Handbooks in Psychology. The Cambridge Handbook of the Learning Sciences*, edited by R. K. Sawyer, 2nd ed., 439–460. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.027.

Artelt, C., and T. Rausch. 2014. "Accuracy of Teacher Judgments: When and for What Reasons?" In *Teachers' Professional Development: Assessment, Training, and Learning*, edited by S. Krolak-Schwerdt, S. Glock, and M. Böhmer, 27–43. Rotterdam: SensePublishers. https://doi.org/10.1007/978-94-6209-536-6_3.

Bauer, E., F. Fischer, J. Kiesewetter, et al. 2020. "Diagnostic Activities and Diagnostic Practices in Medical Education and Teacher Education: An Interdisciplinary Comparison." *Frontiers in Psychology* 11: 1–9. https://doi.org/10.3389/fpsyg.2020.562665.

Bauer, E., M. Greisel, I. Kuznetsov, et al. 2023. "Using Natural Language Processing to Support Peer-Feedback in the Age of Artificial Intelligence: A Cross-Disciplinary Framework and a Research Agenda." *British Journal of Educational Technology* 54: 1–24. https://doi.org/10.1111/bjet.13336.

Bauer, E., M. Sailer, J. Kiesewetter, M. R. Fischer, and F. Fischer. 2022. "Diagnostic Argumentation in Teacher Education: Making the Case for Justification, Disconfirmation, and Transparency." *Frontiers in Education* 7: 1–16. https://doi.org/10.3389/feduc.2022.977631.

Bauer, E., M. Sailer, J. Kiesewetter, et al. 2022. "Learning to Diagnose Students' Behavioral, Developmental, and Learning Disorders in a Simulation-Based Learning Environment for Pre-Service Teachers." In *Learning to Diagnose With Simulations*, edited by F. Fischer and A. Opitz, vol. 91, 97–107. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-89147-3_8.

Bernacki, M. L., M. J. Greene, and N. G. Lobczowski. 2021. "A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)?" *Educational Psychology Review* 33, no. 4: 1675–1715. https://doi.org/10.1007/s10648-021-09615-8.

Bimba, A. T., N. Idris, A. Al-Hunaiyyan, R. B. Mahmud, and N. L. B. M. Shuib. 2017. "Adaptive Feedback in Computer-Based Learning Environments: A Review." *Adaptive Behavior* 25, no. 5: 217–234. https://doi.org/10.1177/1059712317727590.

Black, P., and D. Wiliam. 2009. "Developing the Theory of Formative Assessment." *Educational Assessment, Evaluation and Accountability* 21, no. 1: 5–31. https://doi.org/10.1007/s11092-008-9068-5.

Boshuizen, H. P., H. Gruber, and J. Strasser. 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.

Bradley, E. G., and B. Kendall. 2014. "A Review of Computer Simulations in Teacher Education." *Journal of Educational Technology Systems* 43, no. 1: 3–12. https://doi.org/10.2190/ET.43.1.b.

Butler, A. C., N. Godbole, and E. J. Marsh. 2013. "Explanation Feedback Is Better Than Correct Answer Feedback for Promoting Transfer of Learning." *Journal of Educational Psychology* 105, no. 2: 290–298. https://doi.org/10.1037/a0031026.

Butterfuss, R., R. D. Roscoe, L. K. Allen, K. S. McCarthy, and D. S. McNamara. 2022. "Strategy Uptake in Writing Pal: Adaptive Feedback and Instruction." *Journal of Educational Computing Research* 60, no. 3: 696–721. https://doi.org/10.1177/07356331211045304.

Cavalcanti, A. P., A. Barbosa, R. Carvalho, et al. 2021. "Automatic Feedback in Online Learning Environments: A Systematic Literature Review." *Computers and Education: Artificial Intelligence* 2: 100027. https://doi.org/10.1016/j.caeai.2021.100027.

Chaudhry, M. A., M. Cukurova, and R. Luckin. 2022. "A Transparency Index Framework for AI in Education." In *Lecture Notes in Computer Science. Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks,*

*Practitioners' and Doctoral Consortium*, edited by M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, vol. 13,356, 195–198. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-11647-6_33.

Chen, O., J. C. Castro-Alonso, F. Paas, and J. Sweller. 2018. "Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence From the Spacing Effect." *Educational Psychology Review* 30, no. 2: 483–501. https://doi.org/10.1007/s10648-017-9426-2.

Chernikova, O., N. Heitzmann, M. Stadler, D. Holzberger, T. Seidel, and F. Fischer. 2020. "Simulation-Based Learning in Higher Education: A Meta-Analysis." *Review of Educational Research* 90, no. 4: 499–541. https://doi.org/10.3102/0034654320933544.

Codreanu, E., D. Sommerhoff, S. Huber, S. Ufer, and T. Seidel. 2020. "Between Authenticity and Cognitive Demand: Finding a Balance in Designing a Video-Based Simulation in the Context of Mathematics Teacher Education." *Teaching and Teacher Education* 95: 103146. https://doi.org/10.1016/j.tate.2020.103146.

Colenbrander, D., J. Ricketts, and H. L. Breadmore. 2018. "Early Identification of Dyslexia: Understanding the Issues." *Language, Speech, and Hearing Services in Schools* 49, no. 4: 817–828. https://doi.org/10.1044/2018_LSHSS-DYSLC-18-0007.

Dai, C.-P., and F. Ke. 2022. "Educational Applications of Artificial Intelligence in Simulation-Based Learning: A Systematic Mapping Review." *Computers and Education: Artificial Intelligence* 3: 100087. https://doi.org/10.1016/j.caeai.2022.100087.

Daxenberger, J., A. Csanadi, C. Ghanem, I. Kollar, and I. Gurevych. 2018. "Domain-Specific Aspects of Scientific Reasoning and Argumentation." In *Scientific Reasoning and Argumentation*, edited by F. Fischer, C. A. Chinn, K. Engelmann, and J. Osborne, 34–55. New York, NY: Routledge. https://doi.org/10.4324/9780203731826-3.

De Coninck, K., M. Valcke, I. Ophalvens, and R. Vanderlinde. 2019. "Bridging the Theory-Practice Gap in Teacher Education: The Design and Construction of Simulation-Based Learning Environments." In *Kohärenz in der Lehrerbildung*, edited by K. Hellmann, J. Kreutz, M. Schwichow, and K. Zaki, vol. 5, 263–280. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-23940-4_17.

Dieker, L. A., J. A. Rodriguez, B. Lignugaris-Kraft, M. C. Hynes, and C. E. Hughes. 2014. "The Potential of Simulated Environments in Teacher Education." *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children* 37, no. 1: 21–33. https://doi.org/10.1177/0888406413512683.

Fan, Y., M. Rakovic, J. van der Graaf, et al. 2023. "Towards a Fuller Picture: Triangulation and Integration of the Measurement of Self-Regulated Learning Based on Trace and Think Aloud Data." *Journal of Computer Assisted Learning. Advance Online Publication* 39: 1303–1324. https://doi.org/10.1111/jcal.12801.

Fink, M. C., A. Radkowitsch, E. Bauer, et al. 2021. "Simulation Research and Design: A Dual-Level Framework for Multi-Project Research Programs." *Educational Technology Research and Development* 69, no. 2: 809–841. https://doi.org/10.1007/s11423-020-09876-0.

Fischer, F., E. Bauer, T. Seidel, et al. 2022. "Representational Scaffolding in Digital Simulations—Learning Professional Practices in Higher Education." *Information and Learning Science* 123, no. 11/12: 645–665. https://doi.org/10.1108/ILS-06-2022-0076.

Fischer, M. R. 1999. "CASUS—An Authoring and Learning Tool Supporting Diagnostic Reasoning." *Zeitschrift Für Hochschuldidaktik* 3, no. 1: 87–98. https://scholar.google.de/citations?user=axdugxkaaajandhl=enandoi=sra.

Frerejean, J., J. J. G. van Merriënboer, C. Condron, U. Strauch, and W. Eppich. 2023. "Critical Design Choices in Healthcare Simulation Education: A 4C/ID Perspective on Design That Leads to Transfer." *Advances in Simulation (London, England)* 8, no. 1: 5. https://doi.org/10.1186/s41077-023-00242-7.

Fyfe, E. R., M. S. DeCaro, and B. Rittle-Johnson. 2015. "When Feedback Is Cognitively-Demanding: The Importance of Working Memory Capacity." *Instructional Science* 43, no. 1: 73–91. https://doi.org/10.1007/s11251-014-9323-8.

Gardner, J., M. O'Leary, and L. Yuan. 2021. "Artificial Intelligence in Educational Assessment: Breakthrough? Or Buncombe and Ballyhoo?" *Journal of Computer Assisted Learning* 37, no. 5: 1207–1216. https://doi.org/10.1111/jcal.12577.

Gegenfurtner, A., C. Quesada-Pallarès, and M. Knogler. 2014. "Digital Simulation-Based Training: A Meta-Analysis." *British Journal of Educational Technology* 45, no. 6: 1097–1114. https://doi.org/10.1111/bjet.12188.

Gombert, S., A. Fink, T. Giorgashvili, et al. 2024. "From the Automated Assessment of Student Essay Content to Highly Informative Feedback: A Case Study." *International Journal of Artificial Intelligence in Education. Advance Online Publication* 34: 1378–1416. https://doi.org/10.1007/s40593-023-00387-6.

Grossman, P. 2021. *Teaching Core Practices in Teacher Education*. Cambridge, Massachusetts: Harvard Education Press.

Grossman, P., C. Compton, D. Igra, M. Ronfeldt, E. Shahan, and P. W. Williamson. 2009. "Teaching Practice: A Cross-Professional Perspective." *Teachers College Record: The Voice of Scholarship in Education* 111, no. 9: 2055–2100. https://doi.org/10.1177/016146810911100905.

Habernal, I., and I. Gurevych. 2016. "Which Argument Is More Convincing? Analyzing and Predicting Convincingness of Web Arguments Using Bidirectional LSTM." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, edited by K. Erk and N. A. Smith, vol. 1, 1589–1599. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1150.

Harks, B., K. Rakoczy, J. Hattie, M. Besser, and E. Klieme. 2014. "The Effects of Feedback on Achievement, Interest and Self-Evaluation: The Role of Feedback's Perceived Usefulness." *Educational Psychology* 34, no. 3: 269–290. https://doi.org/10.1080/01443410.2013.785384.

Hatala, R., D. A. Cook, B. Zendejas, S. J. Hamstra, and R. Brydges. 2014. "Feedback for Simulation-Based Procedural Skills Training: A Meta-Analysis and Critical Narrative Synthesis." *Advances in Health Sciences Education: Theory and Practice* 19, no. 2: 251–272. https://doi.org/10.1007/s10459-013-9462-8.

Hattie, J., and H. Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77, no. 1: 81–112. https://doi.org/10.3102/003465430298487.

Heitzmann, N., T. Seidel, A. Hetmanek, et al. 2019. "Facilitating Diagnostic Competences in Simulations in Higher Education A Framework and a Research Agenda." *Frontline Learning Research* 7, no. 4: 1–24. https://doi.org/10.14786/flr.v7i4.384.

Heitzmann, N., M. Stadler, C. Richters, et al. 2023. "Learners' Adjustment Strategies Following Impasses in Simulations—Effects of Prior Knowledge." *Learning and Instruction* 83: 101632. https://doi.org/10.1016/j.learninstruc.2022.101632.

Helleve, I., L. Eide, and M. Ulvik. 2023. "Case-Based Teacher Education Preparing for Diagnostic Judgement." *European Journal of Teacher Education* 46, no. 1: 50–66. https://doi.org/10.1080/02619768.2021.1900112.

Herppich, S., A.-K. Praetorius, N. Förster, et al. 2018. "Teachers' Assessment Competence: Integrating Knowledge-, Process-, and Product-Oriented Approaches Into a Competence-Oriented Conceptual Model." *Teaching and Teacher Education* 76, no. 4: 181–193. https://doi.org/10.1016/j.tate.2017.12.001.

Hetmanek, A., K. Engelmann, A. Opitz, and F. Fischer. 2018. "Beyond Intelligence and Domain Knowledge: Scientific Reasoning and Argumentation as a Set of Cross-Domain Skills." In *Scientific Reasoning*

*and Argumentation*, edited by F. Fischer, C. A. Chinn, K. Engelmann, and J. Osborne, 203–226. New York, NY: Routledge.

Hollis, R. B., and C. A. Was. 2016. "Mind Wandering, Control Failures, and Social Media Distractions in Online Learning." *Learning and Instruction* 42: 104–112. https://doi.org/10.1016/j.learninstruc.2016.01.007.

Ifenthaler, D., C. Schumacher, and J. Kuzilek. 2023. "Investigating Students' Use of Self-Assessments in Higher Education Using Learning Analytics." *Journal of Computer Assisted Learning* 39, no. 1: 255–268. https://doi.org/10.1111/jcal.12744.

Issenberg, S. B., W. C. McGaghie, E. R. Petrusa, D. Lee Gordon, and R. J. Scalese. 2005. "Features and Uses of High-Fidelity Medical Simulations That Lead to Effective Learning: A BEME Systematic Review." *Medical Teacher* 27, no. 1: 10–28. https://doi.org/10.1080/01421590500046924.

Jansen, T., L. Höft, L. Bahr, et al. 2024. "Comparing Generative AI and Expert Feedback to Students' Writing: Insights From Student Teachers." *Psychologie in Erziehung und Unterricht* 71, no. 2: 80–92. https://doi.org/10.2378/peu2024.art08d.

Jossberger, H., J. Breckwoldt, and H. Gruber. 2022. "Promoting Expertise Through Simulation (PETS): A Conceptual Framework." *Learning and Instruction* 82: 101686. https://doi.org/10.1016/j.learninstruc.2022.101686.

Kalyuga, S., and J. L. Plass. 2018. "Cognitive Load as a Local Characteristic of Cognitive Processes: Implications for Measurement Approaches." In *Cognitive Load Measurement and Application*, edited by R. Z. Zheng, 59–74. New York: Routledge. https://psycnet.apa.org/doi/10.4324/9781315296258-5.

Kasneci, E., K. Sessler, S. Küchemann, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kaufman, D., and A. Ireland. 2016. "Enhancing Teacher Education With Simulations." *TechTrends* 60, no. 3: 260–267. https://doi.org/10.1007/s11528-016-0049-0.

Kiesewetter, J., M. Sailer, V. M. Jung, et al. 2020. "Learning Clinical Reasoning: How Virtual Patient Case Format and Prior Knowledge Interact." *BMC Medical Education* 20, no. 73: 1–10. https://doi.org/10.1186/s12909-020-1987-y.

Kolovou, D., A. Naumann, J. Hochweber, and A.-K. Praetorius. 2021. "Content-Specificity of Teachers' Judgment Accuracy Regarding Students' Academic Achievement." *Teaching and Teacher Education* 100, no. 4: 103298. https://doi.org/10.1016/j.tate.2021.103298.

Kosel, C., I. Wolter, and T. Seidel. 2021. "Profiling Secondary School Students in Mathematics and German Language Arts Using Learning-Relevant Cognitive and Motivational-Affective Characteristics." *Learning and Instruction* 73: 101434. https://doi.org/10.1016/j.learninstruc.2020.101434.

Kramer, M., C. Förtsch, T. Seidel, and B. J. Neuhaus. 2021. "Comparing Two Constructs for Describing and Analyzing Teachers' Diagnostic Processes." *Studies in Educational Evaluation* 68, no. 7: 100973. https://doi.org/10.1016/j.stueduc.2020.100973.

Kron, S., D. Sommerhoff, M. Achtner, et al. 2022. "Cognitive and Motivational Person Characteristics as Predictors of Diagnostic Performance: Combined Effects on Pre-Service Teachers' Diagnostic Task Selection and Accuracy." *Journal für Mathematik-Didaktik* 43, no. 1: 135–172. https://doi.org/10.1007/s13138-022-00200-2.

Kron, S., D. Sommerhoff, M. Achtner, and S. Ufer. 2021. "Selecting Mathematical Tasks for Assessing Student's Understanding: Pre-Service Teachers' Sensitivity to and Adaptive Use of Diagnostic Task Potential in Simulated Diagnostic One-To-One Interviews." *Frontiers in Education* 6: 738. https://doi.org/10.3389/feduc.2021.604568.

Kucirkova, N., L. Gerard, and M. C. Linn. 2021. "Designing Personalised Instruction: A Research and Design Framework." *British Journal of Educational Technology* 52, no. 5: 1839–1861. https://doi.org/10.1111/bjet.13119.

Kuklick, L., S. Greiff, and M. A. Lindner. 2023. "Computer-Based Performance Feedback: Effects of Error Message Complexity on Cognitive, Metacognitive, and Motivational Outcomes." *Computers and Education* 200: 104785. https://doi.org/10.1016/j.compedu.2023.104785.

Lachner, A., H. Jarodzka, and M. Nückles. 2016. "What Makes an Expert Teacher? Investigating Teachers' Professional Vision and Discourse Abilities." *Instructional Science* 44, no. 3: 197–203. https://doi.org/10.1007/s11251-016-9376-y.

Lim, L.-A., S. Gentili, A. Pardo, et al. 2021. "What Changes, and for Whom? A Study of the Impact of Learning Analytics-Based Process Feedback in a Large Course." *Learning and Instruction* 72: 202. https://doi.org/10.1016/j.learninstruc.2019.04.003.

Machts, N., O. Chernikova, T. Jansen, M. Weidenbusch, F. Fischer, and J. Möller. 2024. "Categorization of Simulated Diagnostic Situations and the Salience of Diagnostic Information: Conceptual Framework." *Zeitschrift Für Pädagogische Psychologie* 38, no. 1–2: 3–13. https://doi.org/10.1024/1010-0652/a000364.

Machts, N., J. Kaiser, F. T. Schmidt, and J. Möller. 2016. "Accuracy of Teachers' Judgments of Students' Cognitive Abilities: A Meta-Analysis." *Educational Research Review* 19: 85–103. https://doi.org/10.1016/j.edurev.2016.06.003.

Makel, M. C., and J. A. Plucker. 2014. "Facts Are More Important Than Novelty: Replication in the Education Sciences." *Educational Researcher* 43, no. 6: 304–316. https://doi.org/10.3102/0013189X14545513.

Mertens, U., B. Finn, and M. A. Lindner. 2022. "Effects of Computer-Based Feedback on Lower- and Higher-Order Learning Outcomes: A Network Meta-Analysis." *Journal of Educational Psychology* 114, no. 8: 1743–1772. https://doi.org/10.1037/edu0000764.

Nachtigall, V., D. W. Shaffer, and N. Rummel. 2022. "Stirring a Secret Sauce: A Literature Review on the Conditions and Effects of Authentic Learning." *Educational Psychology Review* 34, no. 3: 1479–1516. https://doi.org/10.1007/s10648-022-09676-3.

Narciss, S., C. Prescher, L. Khalifah, and H. Körndle. 2022. "Providing External Feedback and Prompting the Generation of Internal Feedback Fosters Achievement, Strategies and Motivation in Concept Learning." *Learning and Instruction* 82: 101658. https://doi.org/10.1016/j.learninstruc.2022.101658.

Narciss, S., S. Sosnovsky, L. Schnaubert, et al. 2014. "Exploring Feedback and Student Characteristics Relevant for Personalizing Feedback Strategies." *Computers and Education* 71, no. 2: 56–76. https://doi.org/10.1016/j.compedu.2013.09.011.

Nickl, M., S. A. Huber, D. Sommerhoff, E. Codreanu, S. Ufer, and T. Seidel. 2022. "Video-Based Simulations in Teacher Education: The Role of Learner Characteristics as Capacities for Positive Learning Experiences and High Performance." *International Journal of Educational Technology in Higher Education* 19, no. 1: 45. https://doi.org/10.1186/s41239-022-00351-9.

Nicol, D. 2021. "The Power of Internal Feedback: Exploiting Natural Comparison Processes." *Assessment and Evaluation in Higher Education* 46, no. 5: 756–778. https://doi.org/10.1080/02602938.2020.1823314.

Ninaus, M., and M. Sailer. 2022. "Closing the Loop—The Human Role in Artificial Intelligence for Education." *Frontiers in Psychology* 13: 956798. https://doi.org/10.3389/fpsyg.2022.956798.

Norman, G. R., S. D. Monteiro, J. Sherbino, J. S. Ilgen, H. G. Schmidt, and S. Mamede. 2017. "The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking." *Academic Medicine: Journal of the Association of American Medical Colleges* 92, no. 1: 23–30. https://doi.org/10.1097/ACM.0000000000001421.

Norman, G. R., M. Young, and L. Brooks. 2007. "Non-Analytical Models of Clinical Reasoning: The Role of Experience." *Medical Education* 41, no. 12: 1140–1145. https://doi.org/10.1111/j.1365-2923.2007.02914.x.

OpenAI. 2023. "GPT-4 Technical Report." https://arxiv.org/pdf/2303.08774.

Panadero, E., G. T. L. Brown, and J.-W. Strijbos. 2016. "The Future of Student Self-Assessment: A Review of Known Unknowns and Potential Directions." *Educational Psychology Review* 28, no. 4: 803–830. https://doi.org/10.1007/s10648-015-9350-2.

Peltier, T. K., E. K. Washburn, B. C. Heddy, and E. Binks-Cantrell. 2022. "What Do Teachers Know About Dyslexia? It's Complicated!" *Reading and Writing* 35, no. 9: 2077–2107. https://doi.org/10.1007/s11145-022-10264-8.

Pfeiffer, J., C. M. Meyer, C. Schulz, et al. 2019. "Famulus: Interactive Annotation and Feedback Generation for Teaching Diagnostic Reasoning." In *The 2019 EMNLP and IJCNLP Conference—Proceedings of System Demonstrations: November 3–7, 2019, Hong Kong, China*, edited by S. Padó, 73–78. Stroudsburg, PA: Association for Computational Linguistics (ACL). https://aclanthology.org/D19-3013.pdf.

Plass, J. L., and S. Pawar. 2020. "Toward a Taxonomy of Adaptivity for Learning." *Journal of Research on Technology in Education* 52, no. 3: 275–300. https://doi.org/10.1080/15391523.2020.1719943.

Raković, M., S. Iqbal, T. Li, et al. 2023. "Harnessing the Potential of Trace Data and Linguistic Analysis to Predict Learner Performance in a Multi-Text Writing Task." *Journal of Computer Assisted Learning* 39, no. 3: 703–718. https://doi.org/10.1111/jcal.12769.

Sailer, M., E. Bauer, R. Hofmann, et al. 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.

Sarbu-Rothsching, S., C. Plett, J. Steffen, et al. 2022. Effekte Von Feedbackart Und Bearbeitungsform Auf Die Diagnosekompetenz Medizinstudierender Beim Lernen Mit Online-Fallsimulationen—Ergebnisse Einer Experimentellen Feldstudie (Effects of Feedback Type and Social Learning Format on Medical Students' Diagnostic Competence When Learning With Online Case Simulations—Results of an Experimental Field Study) Halle, Germany: Paper Presented at the Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA).

Sauvé, L., L. Renaud, D. Kaufman, and J.-S. Marquis. 2007. "Distinguishing Between Games and Simulations: A Systematic Review." *Journal of Educational Technology and Society* 10, no. 3: 247–256.

Schulz, C., C. M. Meyer, and I. Gurevych. 2019. "Challenges in the Automatic Analysis of Students' Diagnostic Reasoning." In *The 57th Annual Meeting of the Association for Computational Linguistics—Proceedings of the Conference: July 28–August 2, 2019, Florence, Italy*, edited by A. Korhonen, D. Traum, and L. Màrquez, 6974–6981. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.1609/aaai.v33i01.33016974.

Shute, V. J. 2008. "Focus on Formative Feedback." *Review of Educational Research* 78, no. 1: 153–189. https://doi.org/10.3102/0034654307313795.

Stadler, M., M. Sailer, and F. Fischer. 2021. "Knowledge as a Formative Construct: A Good Alpha Is Not Always Better." *New Ideas in Psychology* 60: 100832. https://doi.org/10.1016/j.newideapsych.2020.100832.

Tetzlaff, L., F. Schmiedek, and G. Brod. 2021. "Developing Personalized Education: A Dynamic Framework." *Educational Psychology Review* 33, no. 3: 863–882. https://doi.org/10.1007/s10648-020-09570-w.

Thompson, M., K. Owho-Ovuakporie, K. Robinson, Y. J. Kim, R. Slama, and J. Reich. 2019. "Teacher Moments: A Digital Simulation for Preservice Teachers to Approximate Parent–Teacher Conversations." *Journal of Digital Learning in Teacher Education* 35, no. 3: 144–164. https://doi.org/10.1080/21532974.2019.1587727.

Urhahne, D., and L. Wijnia. 2021. "A Review on the Accuracy of Teacher Judgments." *Educational Research Review* 32, no. 4: 100,374. https://doi.org/10.1016/j.edurev.2020.100374.

van der Graaf, J., L. Lim, Y. Fan, et al. 2022. "The Dynamics Between Self-Regulated Learning and Learning Outcomes: An Exploratory Approach and Implications." *Metacognition and Learning* 17, no. 3: 745–771. https://doi.org/10.1007/s11409-022-09308-9.

Van Schoors, R., J. Elen, A. Raes, S. Vanbecelaere, and F. Depaepe. 2023. "The Charm or Chasm of Digital Personalized Learning in Education: Teachers' Reported Use." *Perceptions and Expectations. TechTrends* 67, no. 2: 315–330. https://doi.org/10.1007/s11528-022-00802-0.

Wisniewski, B., K. Zierer, and J. Hattie. 2020. "The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research." *Frontiers in Psychology* 10: 1–14. https://doi.org/10.3389/fpsyg.2019.03087.

Zheng, L., L. Zhong, and J. Niu. 2021. "Effects of Personalised Feedback Approach on Knowledge Building, Emotions, Co-Regulated Behavioural Patterns and Cognitive Load in Online Collaborative Learning." *Assessment and Evaluation in Higher Education* 47, no. 1: 109–125. https://doi.org/10.1080/02602938.2021.1883549.

Zhu, M., O. L. Liu, and H.-S. Lee. 2020. "The Effect of Automated Feedback on Revision Behavior and Learning Gains in Formative Assessment of Scientific Argument Writing." *Computers and Education* 143, no. 2: 103668. https://doi.org/10.1016/j.compedu.2019.103668.

## Appendix A

Description of one example case from the simulation environment (all case materials are publicly available in an online repository accessible at https://osf.io/hn7wm/):

At the beginning of the simulation, the participating preservice teachers were asked to put themselves in the role of a teacher who is teaching a class with a student who has noticeable performance-related problems, which may potentially be clinically relevant (e.g., reading and writing disorder). One example is the case of first-grader Anton.

The initial problem statement for the case describes Anton as a 1st-grade primary school student, 7 years old, who shows significant problems in the school subject of German, although he does well in other subjects.

Socially, Anton is an outgoing boy with many friends and enjoys activities such as hide-and-seek, climbing and tagging. His academic challenges are most noticeable in his reading and writing. Although he is familiar with the German language, he struggles to choose the correct letters or forgets letters when writing. During dictation, his pace slows over time, and his handwriting becomes less legible. He frequently mixes up letters or fails to apply capitalisation and spelling rules consistently, even with words he has previously spelled correctly. Reading also poses a problem for Anton. He is slow and tends to skip parts of words or entire words, particularly with unfamiliar vocabulary. After reading, he cannot summarise or answer questions about the text. An exercise involving syllable clapping further highlighted his difficulties in identifying the number of syllables or letters in a word.

A discussion with the crafts teacher reveals no concerns about the crafts subjects. However, a conversation with Anton himself shows a self-aware yet struggling student. He acknowledges his difficulties in German, feeling that others do well where he cannot, and is particularly self-conscious about reading aloud. Anton's parents, in a meeting, expressed their contentment with his overall positive experiences during the first school year and spoke about his diligent homework habits. However, they also note his struggles with German assignments, where he repeatedly makes the same mistakes, despite understanding the corrections when reviewed with his mother. His slow reading speed and an initial reluctance toward the activity are confirmed by his parents as well.

From the observations, available materials (see Figure 1), and conversation protocols, while Anton appears to be an emotionally well-adjusted and socially integrated student, his reading and writing issues are evident. Overall, the case information was designed in a way that the diagnosis of a reading and writing disorder is the most likely clinical

diagnosis. Yet, both adaptive and static feedback for this case caution against making an official diagnosis of a reading and spelling disorder at the end of the first grade due to the natural variances in learning development at this stage. It is recommended to continue monitoring Anton's progress into the second grade and, if problems persist, to involve a school psychologist for standardised testing to make a definitive assessment.

## Appendix B

Coding scheme for rating the justification quality of participants' explanations addressing the example case described above:

The following coding categories (see Table B1) represent the six primary pieces of evidence for the diagnosis considered accurate for this case (reading and writing disorder). For each of the following coding categories, a maximum of one point is assigned per student response (if included in the explanation) or no point is assigned (if not in the explanation).

**TABLE B1** | Coding scheme for rating the justification quality of participants' explanations addressing the simulated case 'Anton'.

| Evidence (coding category) | Example information from the case |
|---|---|
| 1. Low reading speed | He reads much slower than his classmates. |
| 2. Low reading comprehension | He cannot understand the meaning of the text.<br>He answers questions about the text only based on general knowledge. |
| 3. Low reading accuracy | He reads 'ie' instead of 'ei'. |
| 4. Spelling difficulties | He makes many spelling mistakes in dictation exercises.<br>He frequently confuses letters. |
| 5. No significant performance problems in other subjects | He has no difficulties in math.<br>He has no noticeable problems in the other subjects. |
| 6. Problematic phonological awareness | He has problems with identifying syllables.<br>He did not like music lessons even in preschool.<br>He has difficulties getting the rhythm when clapping syllables. |