Learning and Instruction 94 (2024) 101994

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Effects of real-time adaptivity of scaffolding: Supporting pre-service mathematics teachers' assessment skills in simulations

Michael Nickl[a,b,*], Daniel Sommerhoff[b], Anika Radkowitsch[b], Sina A. Huber[c], Elisabeth Bauer[a,d], Stefan Ufer[e], Jan L. Plass[f], Tina Seidel[a,g]

[a] TUM School of Social Sciences and Technology, Technical University of Munich (TUM), Arcisstr. 21, 80333, Munich, Germany
[b] Department of Mathematics Education, IPN – Leibniz Institute for Science and Mathematics Education, Olshausenstr. 62, 24118, Kiel, Germany
[c] Technische Hochschule Ingolstadt (THI), Esplanade 10, 85049, Ingolstadt, Germany
[d] Learning Analytics and Educational Data Mining, University of Augsburg, Universitätsstr. 10, 86159, Augsburg, Germany
[e] Department of Mathematics, Ludwig-Maximilians-Universität München (LMU), Theresienstr. 39, 80333, Munich, Germany
[f] New York University (NYU), 196 Mercer St., New York, NY, 10012-1126, USA
[g] TUM School of Medicine and Health, Technical University of Munich (TUM), Ismaninger Straße. 22, 81675, Munich, Germany

ARTICLE INFO

Keywords:
Teacher assessment skills
Mathematical proof skills
Adaptivity
Scaffolding
Real-time process analysis
Pre-service teacher training

ABSTRACT

Background: Scaffolding pre-service teachers' assessment process in video-based simulations can enhance their acquisition and refinement of assessment skills, for example, needed for accurate judgments of students' mathematical proof skills. Adapting this scaffolding to learners' individual learning processes, for example, based on text data during the assessment process, brings potential for increased learning gains.
Aims: In this study, we investigated the effectiveness of adaptive scaffolding based on real-time process data, specifically targeting pre-service mathematics teachers' assessment skills regarding students' mathematical proof skills in geometry.
Sample: Participants were 245 pre-service teachers.
Methods: In a pre- and post-test, participants completed a video-based simulation to measure their assessment skills regarding students' mathematical proof skills. During the intervention, participants were randomly assigned to complete the video-based simulation (i) without scaffolding, (ii) with non-adaptive scaffolding, or (iii) with adaptive scaffolding.
Results: We did not find significant benefits of adaptive scaffolding in enhancing pre-service teachers' judgment accuracy, aligning with prior research. For an in-depth analysis, we developed and applied a scheme to systematically validate design decisions for adaptive support. This scheme focuses on the selection and measurement of the source of adaptation and the employed support mechanisms. Applying this scheme pointed towards effects of adaptive scaffolding during the assessment process.
Conclusions: This study highlights the need for proximal measures to describe learning in short interventions, explores the intricacies of adaptive scaffolding, such as overlapping with design-loop adaptivity or the accuracy of automated coding, and provides a scheme for an in-depth evaluation of the adaptivity of scaffolding.

1. Introduction

Student assessment is a crucial task in teachers' professional lives. Proficiency in accurately assessing students' individual knowledge and skills is often associated with higher learning gains of the students (Leuders, Loibl, Sommerhoff, Herppich, & Praetorius, 2022). However, the meta-analysis of Südkamp, Kaiser, and Möller (2012) indicates a need for enhancing teachers' assessment skills, also highlighting the importance of authentic training opportunities in university teacher education for effective skill transfer into later practice (Grossman et al., 2009).

To this end, simulations are increasingly used in higher education for acquiring complex skills, including teachers' assessment skills. For instance, the Simulated Classroom (Südkamp, Möller, & Pohlmann,

* Corresponding author. Arcisstr. 21, 80333, Munich, Germany.
E-mail addresses: michael.nickl@tum.de (M. Nickl), sommerhoff@leibniz-ipn.de (D. Sommerhoff), radkowitsch@leibniz-ipn.de (A. Radkowitsch), sina.huber@thi.de (S.A. Huber), elisabeth.bauer@uni-a.de (E. Bauer), jan.plass@nyu.edu (J.L. Plass), tina.seidel@tum.de (T. Seidel).

https://doi.org/10.1016/j.learninstruc.2024.101994
Received 31 January 2024; Received in revised form 23 July 2024; Accepted 9 August 2024
Available online 20 August 2024

2008) requires pre-service teachers to assess students' knowledge. In the Simulated Classroom, pre-service teachers simultaneously assess ten students' mathematical knowledge by choosing mathematical tasks from a task set and selecting students to solve the mathematical task (see Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013, for a medical education simulation). Enriching simulations in higher education, such as the Simulated Classroom, with additional support such as scaffolding has shown promise to further enhance learning gains (Belland, Walker, Kim, & Lefler, 2017; Chernikova, Heitzmann, Fink, et al., 2020). For example, conceptual prompts as scaffolds have been positively evaluated in simulations targeting pre-service teachers' assessment skills (Schons, Obersteiner, Reinhold, Fischer, & Reiss, 2022; Sommerhoff, Codreanu, Nickl, Ufer, & Seidel, 2023). Given the heterogeneity of pre-service teachers' prerequisites for learning assessment skills (Nickl, Sommerhoff, Codreanu, Ufer, Seidel, 2023), adapting scaffolding to individual needs appears a promising yet underexplored avenue.

Research from the field of self-regulated learning has started to explore and evaluate scaffolding that adapts to learners' needs (Chou, Lai, Chao, Tseng, & Liao, 2018; Su, 2020). For example, in a study by Lim et al. (2023), university students used a hypermedia learning tool and received scaffolding that encouraged them to engage in previously unattempted self-regulated learning activities and to explore webpages they had not yet visited.

More generally, when scaffolding is adaptive, the decision on how to scaffold is typically based on the extent of learner-related variable(s) as the *source of adaptation* (Vandewaetere, Desmet, & Clarebout, 2011). Its measurement requires balancing a valid analysis of the learning process against minimal intrusion into this learning process (Plass & Pawar, 2020). For instance, analyzing texts from learners' notes can offer a non-intrusive way to understand their learning processes but may be time-consuming when manually coded (Aleven, McLaughlin, Glenn, & Koedinger, 2017; Shute & Zapata-Rivera, 2012). Recent innovative approaches offer real-time text data analysis, which showed potential for adaptive feedback (Cavalcanti et al., 2021). Yet, whether the application of these approaches can reach the expected potential for scaffolding in the context of teachers' assessment skills is largely uncharted.

Our study seeks to address this gap. Amidst limited and mixed findings on adaptive scaffolding (Belland et al., 2017), we evaluate the effectiveness of adaptive scaffolding based on real-time text analysis compared to non-adaptive scaffolding and no scaffolding in the context of pre-service mathematics teachers' assessment skills regarding mathematical proof skills, including an in-depth analysis of the most relevant design decisions in the development process of adaptive scaffolding.

2. Theoretical background

2.1. Mathematics teachers' assessment skills

We define assessment skills as the teachers' ability and knowledge to accurately assess student characteristics (Urhahne & Wijnia, 2021). The concept overlaps with other related concepts such as diagnostic competences (Heitzmann et al., 2019) or noticing skills (van Es & Sherin, 2002), all integral to teachers' professional competence related to student assessment (Leuders et al., 2022).

Across domains, a key measure of teachers' assessment skills is their judgment accuracy (Urhahne & Wijnia, 2021), by which prior research has highlighted the need for an improvement of teachers' assessment skills (Südkamp et al., 2012). Despite its status as a standard measure, its predictive power on student achievement is debated, with process measures potentially being more indicative (Gabriele, Joram, & Park, 2016). Consequently, recent research has started exploring the assessment process (Brandl, Richters, Radkowitsch, Obersteiner, & Stadler, 2021; Herppich et al., 2018), offering deeper insights into the acquisition of assessment skills (Heitzmann et al., 2019). From an information processing view, the assessment process involves the *noticing* of relevant (observable) classroom events, so-called cues, and a meaningful *interpretation* drawing on the teachers' knowledge, particularly their pedagogical content knowledge (PCK; Meschede, Fiebranz, Möller, & Steffensky, 2017; van Es & Sherin, 2002). To investigate both aspects, the assessment process can be examined through text data analysis (Codreanu, Sommerhoff, Huber, Ufer, & Seidel, 2021). This allows insights into individual assessment processes, for example, whether teachers notice important cues for the assessed student characteristic and interpret these cues using relevant knowledge (Herppich et al., 2018).

For instance, relevant knowledge encompasses in-depth knowledge about indicators, that is, sub-constructs that represent proximal components of the assessed characteristic (e.g., the need for autonomy as an indicator of intrinsic motivation). For accurate judgment, covering all relevant indicators during the assessment process is crucial (Wyatt-Smith & Klenowski, 2013). Lower coverage of these indicators may prevent teachers from comprehensively assessing the targeted student characteristic. Conversely, knowing about relevant indicators may help teachers to notice more cues regarding each indicator, possibly leading to higher coverage during the assessment process and improved judgment accuracy (Brunswik, 1955; Urhahne & Wijnia, 2021). Thus, the nature of assessment processes may vary across different assessment situations due to different domain-specific indicators aligning with the widely acknowledged domain-specificity of assessment skills (Spinath, 2005).

To accurately assess a complex mathematical skill such as mathematical proving, a set of indicators needs to be covered: mathematical content knowledge, methodological knowledge, and problem-solving strategies (Chinnappan, Ekanayake, & Brown, 2012). Mathematical content knowledge refers to the students' knowledge about mathematical definitions, theorems, and properties of mathematical objects (Weigand et al., 2014). Methodological knowledge encompasses students' knowledge about the concept of mathematical proof itself, for example, which types of arguments are allowed in mathematical proofs (Sporn, 2023). Lastly, problem-solving strategies involve knowledge about the students' heuristic strategies in the context of the mathematical proof and monitoring strategies for the proving process (Schoenfeld, 1992). However, these indicators are not equally easy assessable: empirical evidence shows that pre-service teachers' judgment accuracy is higher for mathematical content knowledge than for methodological knowledge, with the least accuracy in problem-solving strategies, but also with room for improvement regarding all indicators (Nickl, Sommerhoff, Codreanu, et al., 2023).

2.2. Video-based simulations & scaffolding

During induction to school practice after initial university training, novice teachers often encounter significant challenges when facing the complexities of classroom realities, a phenomenon frequently referred to as the "practice shock" (Stokking, Leenders, Jong, & van Tartwijk, 2003). This situation underscores the importance of providing realistic and interactive training opportunities for pre-service teachers to acquire and improve their assessment skills at university. Simulations, in particular, offer a promising avenue for this kind of practical training (Heitzmann et al., 2019).

Simulations function as approximations of practice, mirroring real teaching scenarios (Grossman et al., 2009). These interactive environments enable pre-service teachers to engage with scenarios reflective of professional practice, focusing particularly on developing complex skills such as assessment skills (Cook et al., 2013). Making these simulations computer-based and incorporating video into the simulations brings additional advantages. Besides the scalability of the resulting simulation as compared to, for example, role-play simulations, video-based simulations can preserve the authenticity of teaching scenarios (Gaudin & Chaliès, 2015). Additionally, video-based simulations allow the segmentation of complex classroom realities into manageable learning units by using authentic scripted videos (Böttcher & Thiel, 2018; Codreanu et al., 2021). For instance, the *DiKoBi* simulation employs scripted video excerpts from 5th-grade biology lessons, where pre-service teachers are tasked with assessing the quality of biology-specific instruction. The use of these scripted videos allows the simulation to focus on biology-specific instructional quality as a distinct learning objective (Kramer et al., 2020).

Beginner learners, such as pre-service teachers, may still face challenges in dealing with the complex tasks in such video-based simulations (Schons et al., 2022). To support them, scaffolding is widely recognized as effective for enhancing current performance during and future performance after its provision (Belland et al., 2017; Hardy, Decristan, & Klieme, 2019). Originally introduced by Wood, Bruner, and Ross (1976), scaffolding has evolved to encompass various strategies for supporting learners in achieving goals beyond their unassisted efforts. In simulations targeting assessment skills, scaffolding in the form of prompts has shown promise for increasing both current scaffolded and future non-scaffolded performance, as indicated by a meta-analysis of Chernikova, Heitzmann, Fink, et al. (2020) with an effect size of $g = 0.47$ for prompts compared to $g = 0.26$ for no prompts. Prompts can range from general questions to specific instructions, guiding learners' activities within a learning environment (Bannert, 2009). For instance, content-specific prompts that activate pre-service teachers' PCK have been effective in enhancing current assessment performance and future non-scaffolded assessment performance (Schons et al., 2022). As a second example, the prompts in the video-based simulation of Sommerhoff et al. (2023) serve two purposes within the assessment process: First, these prompts asked pre-service teachers to pay special attention to a certain cue in the upcoming video, such as a student's explanation of a mathematical property. By doing so, the prompts support the *noticing* of cues in the videos by providing relevant events in the upcoming video. Second, these prompts asked pre-service teachers to evaluate the provided cue, explicitly naming one indicator of mathematical proof skills that the cue helps to assess. This explicit mention of the indicator in the prompt activates relevant PCK, thereby supporting a knowledge-based *interpretation* of the cues. Both aspects are theoretically important to teachers' assessment skills (van Es & Sherin, 2002), and the empirical effectiveness of such targeted content-specific prompts has been demonstrated (Sommerhoff et al., 2023).

In summary, video-based simulations provide an effective platform for enhancing pre-service teachers' assessment skills. By mimicking real teaching tasks and stimulating active participation in a complexity-reduced setting, these simulations address the gap between theoretical knowledge and practical application, which is crucial for pre-service teachers. Scaffolding, particularly well-designed prompts, further support pre-service teachers during their assessment process in such simulations, making them a valuable tool in facilitating teachers' assessment skills.

2.3. *Adaptive scaffolding and adaptivity*

Given pre-service teachers' diverse prerequisites for acquiring assessment skills (Pickal et al., 2023), one-size-fits-all scaffolding may not fit every pre-service teacher's zone of proximal development (ZPD; Vygotsky, 1978), and an adaptation of scaffolding to the pre-service teachers' ZPDs is promising. For example, Nickl, Sommerhoff, Codreanu, et al. (2023) found that pre-service teachers with high Content Knowledge (CK) and PCK are already comparably accurate in assessing students' methodological knowledge compared to those with lower CK and PCK, but they still face challenges in accurately assessing students' problem-solving strategies, suggesting benefits of a thoughtful adaptation of scaffolding to each pre-service teacher's ZPD.

In traditional classroom settings, scaffolding is characterized by its contingency, transfer of responsibility, and fading (van de Pol, Volman, & Beishuizen, 2010; Wood et al., 1976). This intrinsically implies adaptivity, which ensures the continuous alignment of the support with the learners' current needs, gradually transferring responsibility and eventually removing support as competence increases (Hardy et al., 2019). The alignment of support with learners' needs is expected not only to enhance performance during learning tasks but also to ensure that learners work within their ZPD, ensuring sustained learning gains even after the adaptive support fades (Vygotsky, 1978). The adaptation to learners' needs, crucial in classroom settings (Corno, 2008; Hardy et al., 2019), remains notably underutilized in computer-based learning environments (Belland et al., 2017). High technological demands (e.g., Pfeiffer et al., 2019) may complicate adaptation in real-time (Belland, 2014), though its effectiveness in classrooms suggests potential benefits for computer-based scaffolding (van de Pol et al., 2010; Yelland & Masters, 2007).

In their meta-analysis, Belland et al. (2017) found that only 18.9% of studies in computer-based settings reported scaffolding that adapts dynamically to learners' performance, highlighting a discrepancy between the technological capabilities for individualization and its actual implementation. Surprisingly, they found no significant advantage of adaptive over static scaffolding ($g = 0.47$ vs. $g = 0.45$). While the authors suggested power issues might contribute to this reduced effectiveness, a thorough investigation into the underlying causes is warranted. Recognizing the lack of a systematic procedure in prior research that allows an exploration of the limited efficacy of the adaptivity of scaffolding, we propose a novel scheme to facilitate the evaluation of adaptive scaffolding.

2.4. *Scheme for designing adaptivity and present study*

Meta-research on adaptivity in computer-based learning environments often centers around key design questions and outlines both empirical (Aleven et al., 2017; Van Schoors, Elen, Raes, & Depaepe, 2021; Vandewaetere et al., 2011) and theoretically possible (Plass & Pawar, 2020) design solutions to these questions. There is broad agreement on the two design questions 'what (learner variables) to adapt to' and 'how and when to adapt' (Aleven et al., 2017).

Regarding the first question, the specific construct or set of constructs that determine when and which support is provided, is referred to as source of adaptation (Vandewaetere et al., 2011; see Nakic, Granic, & Glavinic, 2015, for an empirical overview). In this regard, Plass and Pawar (2020) highlight an additional practical aspect critical to designing adaptive learning environments: adequately measuring the source of adaptation, which also highly depends on the implemented technological tools (Kardan, Aziz, & Shahpasand, 2015). The second question can be subdivided into 'when to adapt,' 'what to adapt,' and 'how to adapt' in the specific learning environment (Van Schoors et al., 2021; Vandewaetere & Clarebout, 2014), which we summarize under the term "support mechanism."

As guidance for practical design questions in adaptive learning environments, our scheme focuses on the source of adaptation, measuring the source of adaptation, and the support mechanism. The remainder of this section details these categories and their application in our study.

Source of Adaptation. The choice of a source of adaptation typically involves filtering the broad range of relevant learner variables (Vandewaetere et al., 2011), also informed by the feasibility of their measurement. Once selected, it needs to be ensured that the source of adaptation is meaningfully related to the desired outcome and exhibits sufficient variability to reasonably adapt (Shute & Zapata-Rivera, 2012). For instance, a variable unlinked to the learning outcome, directly or indirectly, cannot validly determine beneficial support. Conversely, a variable lacking variance results in uniform support, negating adaptivity's need. 'No variance' is an extreme case; yet, what constitutes 'enough' variance is rarely addressed in literature and remains unclear.

In the present study, we explore the potential of adaptive scaffolding in the context of teachers' assessment skills focusing on mathematical proof skills for facilitating pre-service teachers' judgment accuracy as the desired learning outcome. We rely on the coverage of the three implemented indicators of mathematical proof skills (mathematical content knowledge, methodological knowledge, and problem-solving strategies) as the source of adaptation, as this was related to judgment accuracy and different learner groups with different levels of coverage could be identified in prior research (Nickl, Sommerhoff, Codreanu, et al., 2023).

Measuring the Source of Adaptation. Once selected, the focus shifts to the specific measurement of the chosen source of adaptation (Plass & Pawar, 2020). This requires a balance between reliability and validity and non-intrusiveness. While self-report scales are pragmatic and (mostly) reliable, they are intrusive, interrupt the learning process if continuously measured, and reduce authenticity and immersion (Shute & Zapata-Rivera, 2012). Conversely, less intrusive methods like log data analysis, text data analysis, or eye-tracking need to ensure the reliability and validity (Gašević, Dawson, & Siemens, 2015; Zawacki-Richter, Marín, Bond, & Gouverneur, 2019).

In the present study, the coverage is measured using real-time text data analysis of pre-service teachers' notes to minimize intrusiveness. Automated coding is based on manual coding in prior studies to ensure validity.

Support Mechanism. While classroom teachers use implicit heuristics for their support mechanisms (Corno, 2008; Herppich et al., 2018), computer-based scaffolding demands explicit and systematic implementation of these heuristics (Vandewaetere et al., 2011). These heuristics involve selecting the appropriate time interval for adaptations ('when to adapt'), selecting the support to be adapted ('what to adapt'), and devising an adaptation strategy to determine the specific support provided based on the measured values of the source of adaptation ('how to adapt'). In the context of adaptive scaffolding, the need for contingency implies a short measurement interval, refining the support mechanism to focus on selecting scaffolds that can effectively foster performance and learning in the learning environment and orchestrating them within an adaptation strategy. Although various empirically evaluated scaffolds are available (Belland et al., 2017) and different approaches for adaptation strategies exist (see Vandewaetere et al., 2011), specific decisions, like setting cut-off values for support, often rely on heuristics due to their unique nature and limited research precedence (Shute, 1995).

In the present study, we use positively evaluated conceptual prompts that have increased pre-service teachers' judgment accuracy in prior studies. The adaptation strategy, for which we provide prompts for the easiest assessable but overlooked indicators, is anticipated to enhance scaffolding effectiveness (Aleven et al., 2017; Belland et al., 2017).

3. Research questions

In the present study, we investigated the effectiveness of adaptive scaffolding to facilitate pre-service teachers' skills in assessing students' mathematical proof skills in a video-based simulation. We examined the following research question:

How does judgment accuracy differ among pre-service teachers who receive adaptive scaffolding, non-adaptive scaffolding, or no scaffolding?

We hypothesized that pre-service teachers in the scaffolded conditions reached higher judgment accuracy in the intervention (1a) and the post-test (1b) than those who did not receive the scaffolding (control condition). We expected pre-service teachers receiving adaptive support to reach higher judgment accuracy in the intervention (2a) and the post-test (2b) than those receiving non-adaptive support.

4. Methodology

To test these hypotheses, we conducted an experimental study with a pre-post-test design with three conditions (no scaffolding, non-adaptive scaffolding, adaptive scaffolding). The no scaffolding condition served as the control condition. This study was preregistered (https://osf.io/gk58d/?view_only=8db6adc208e947c39ad6966c994e7791).

4.1. Sample

A total of 245 pre-service teachers from eight German universities participated in the study to avoid university and location specific effects. Among the participants, 100 self-reported as male and 139 as female, while 6 did not disclose their gender. The participants were heterogeneously distributed across various stages of their studies in teacher education (participants' semester: $M = 4.4$, $SD = 3.5$; participants' age: $M = 22.6$, $SD = 3.0$; see electronic supplement 1 for the detailed distribution). They were randomly assigned to control ($N = 85$), non-adaptive scaffolding ($N = 81$), and adaptive scaffolding ($N = 79$) conditions, which also resulted in a similar distribution of gender, age, and semester within these conditions (see electronic supplement 1). This research builds on a previous study that employed the same design. In the previous study, participants were assigned to one of four conditions: control, non-adaptive prompts, motivational intervention, and non-adaptive prompts + motivational intervention. The control ($N = 47$) and non-adaptive prompts ($N = 42$) conditions from that study are identical to the no scaffolding and non-adaptive scaffolding conditions in the current study, respectively. This allowed us to strengthen the comparison with the adaptive scaffolding condition by including the relevant participants in the present study. The results from the prior study, which lacked an adaptive scaffolding condition, cannot answer the present research question. However, they showed intraindividual improvements in participants' judgment accuracy from the pre-test to intervention (significant intraindividual improvement for non-adaptive prompt condition, non-significant improvement for control condition) and suggested the differential effectiveness of non-adaptive prompts (see Nickl, Sommerhoff, Böheim, Ufer, & Seidel, 2023). For data collection in the present study, we aimed (i) to ensure equal-sized conditions and (ii) to achieve a power of at least 0.80 based on a priori power analysis using G*Power, assuming a medium effect size. Data collection was conducted using the Unipark online survey system. Participants were recruited in university teacher education courses without semester or subject restrictions. In some cases, the simulation was integrated into seminars; elsewhere, it was provided as optional supplementary material. Participants consented voluntarily to the use of their data and received a €50 compensation. The data collection methodology was approved by the data protection office of the first author's university.

4.2. Study design

This study comprised a pre-test session (90 min in total, with 30 min assigned for assessing participants' knowledge, 30 min for the video-based simulation, and 30 min for measuring motivational-affective characteristics) and a session including the intervention along with a subsequent post-test (90 min in total, with 30 min assigned for the video-based simulation in the intervention, 30 min for the video-based simulation in the post-test, and 30 min for measuring motivational-affective characteristics and wrap-up). The intervention and post-test session took place at least four days after the pre-test to reduce re-test effects.

4.2.1. Video-based simulation

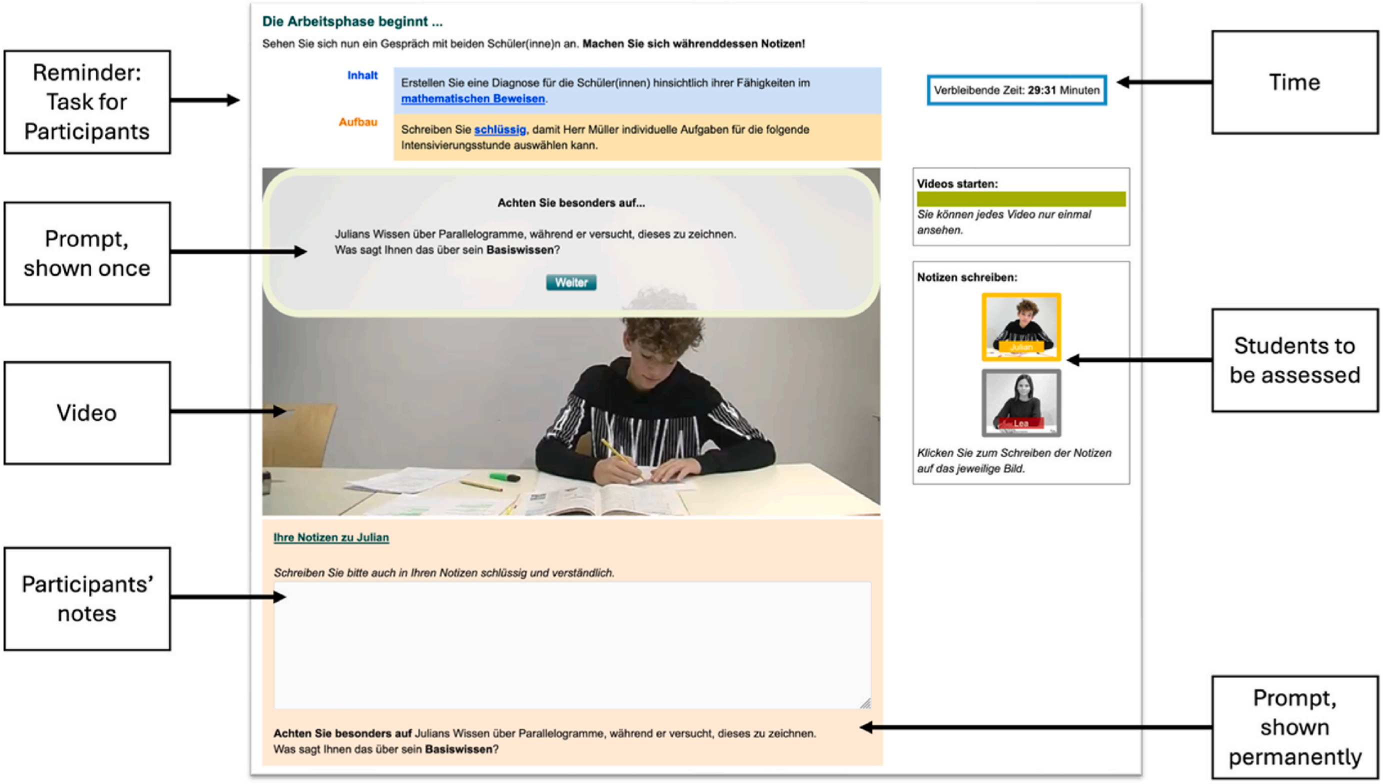
Participants engaged in a video-based simulation during the pre-test, intervention, and post-test, where they assessed the mathematical proof skills of two simulated students by watching videos (see Fig. 1). The simulation addresses assessment skills, a key focus in German teacher education (Kultusministerkonferenz, 2004). Within the introduction to the simulation, participants are asked to imagine themselves in a school practicum, observing an in-service teacher's lesson, a typical component of German teacher education (Arnold et al., 2014). In the simulated lesson, students worked individually on the same mathematical proof task (i.e., to prove that the opposite sides of a parallelogram are equal in length). Each 1-min video depicted a student working on the task and interacting with the teacher. Key diagnostic information included the students' verbal reasoning with the teacher (e.g., 'it is sufficient to validate the assertion only for this parallelogram') and their drawings (e.g., incorrectly drawing a parallelogram) presented in the videos. In total, eight videos were available per student, which showed different stages of the student's solution process in the lesson. When designing the videos, we aimed to include cues for each of the indicators of mathematical proof skills in every video. After being familiarized with the assessment situation and the students' task, participants entered the simulation with the task of assessing two students' mathematical proof skills. During the videos, participants were encouraged to take notes, which were shown throughout all videos of the respective student, and afterward judged the students' skills both in free-text and Likert-scale formats.

In the pre- and post-tests, the same materials were used: participants assessed the same two simulated students (Barbara and Christian) using the same videos. In the pre- and post-test, participants could choose how many videos to watch to assess the students' mathematical proof skills accurately (maximum: ten videos for assessing both students). Letting participants regulate their assessment process by choosing the number of videos mirrors real classroom environments, which also require evaluating the need for additional diagnostic information and, thus, allowing for an authentic measurement of assessment skills (Herppich et al., 2018). During the intervention, they evaluated two different simulated students (Andrew and Doris), watching a fixed number of four videos for each student to ensure equal time-on-task during the intervention.

4.2.2. *Scaffolding*

Pre- and post-test were identical over all conditions. During the intervention, participants in both scaffolding conditions received additional scaffolds integrated into the simulation. Depending on the condition, they received non-adaptive or adaptive scaffolding. We utilized conceptual prompts as scaffolds, presented before each video, targeting one of the three indicators of mathematical proof skills. Thus, the actual intervention occurred during the approximately 25-min period when participants engaged with the prompted videos in the simulation. Following the prompts of Sommerhoff et al. (2023), a typical prompt asked participants to focus on a specific event in the video and draw conclusions about an indicator, such as the student's mathematical content knowledge. An example prompt is "Please pay special attention to the student's knowledge of parallelograms while he draws it. What can you conclude about the student's mathematical content knowledge?" These prompts can be regarded as validated since they were already used in prior studies and showed promise for facilitating mathematics pre-service teachers' assessment skills in these studies (Nickl, Sommerhoff, Böheim, et al., 2023; Sommerhoff et al., 2023).

In the non-adaptive condition, prompts highlighted the most relevant indicator in a video (in total, four prompts on mathematical content knowledge, four prompts on methodological knowledge, and no prompts on problem-solving strategies, see electronic supplement 4), irrespective of the participant's individual prior assessment process. Conversely, in the adaptive condition, prompts for a video were selected based on the coverage of the three indicators of mathematical proof skills within the participant's notes from the previous video (except the first video: prompts as in the non-adaptive condition); thus, per participant, six prompts were adapted in total (4 videos for each of the two students, minus the first video). The adaptive prompts focused on the easiest assessable indicator not yet covered in the participant's notes. For example, if a participant's previous notes did not mention mathematical content knowledge, the following prompt would address it (see Fig. 2). If all indicators were covered in the notes, a laudatory prompt was displayed ("Based on our analyses, you considered all relevant indicators of mathematical proof skills in the last video. Keep it up!"). It's important to note that two videos featuring student Andrew did not include cues about either methodological knowledge or mathematical content knowledge, leading to the respective indicator being skipped in



Die Arbeitsphase beginnt ...

Sehen Sie sich nun ein Gespräch mit beiden Schüler(inne)n an. **Machen Sie sich währenddessen Notizen!**

Erstellen Sie eine Diagnose für die Schüler(innen) hinsichtlich ihrer Fähigkeiten im mathematischen Beweisen.

Schreiben Sie schlüssig, damit Herr Müller individuelle Aufgaben für die folgende Intensivierungsstunde auswählen kann.

Verbleibende Zeit: **29:31** Minuten

Achten Sie besonders auf...

Julians Wissen über Parallelogramme, während er versucht, dieses zu zeichnen.
Was sagt Ihnen das über sein **Basiswissen**?

Videos starten:

Sie können jedes Video nur einmal ansehen.

Notizen schreiben:

Klicken Sie zum Schreiben der Notizen auf das jeweilige Bild.

Ihre Notizen zu Julian

Schreiben Sie bitte auch in Ihren Notizen schlüssig und verständlich.

Achten Sie besonders auf Julians Wissen über Parallelogramme, während er versucht, dieses zu zeichnen.
Was sagt Ihnen das über sein **Basiswissen**?

Fig. 1. Main Screen in the Simulation
Note. Prompts are only displayed for scaffolding conditions during the intervention.

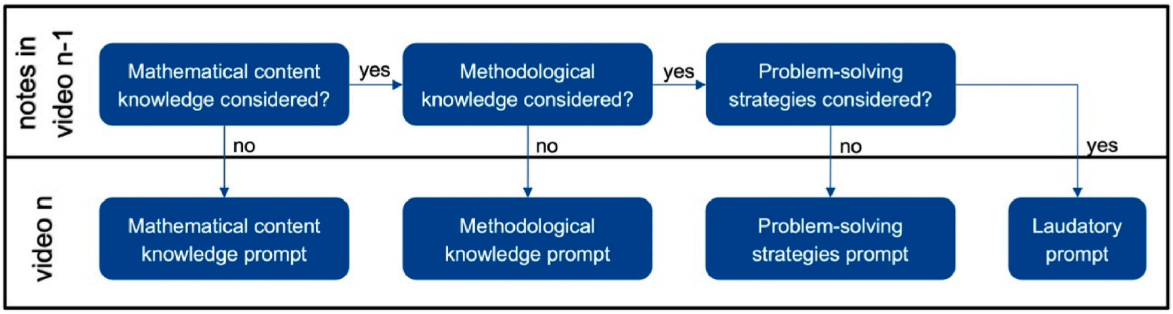

notes in video n-1
video n
Mathematical content knowledge considered?
yes
Methodological knowledge considered?
yes
Problem-solving strategies considered?
no
no
no
yes
Mathematical content knowledge prompt
Methodological knowledge prompt
Problem-solving strategies prompt
Laudatory prompt

Fig. 2. Prompting scheme of the adaptation strategy.

the prompting scheme for those and subsequent videos (Fig. 2). To detect whether the indicators are covered in the participants' notes, we employed a rule-based automated coding procedure based on a naïve Bayes approach (see electronic supplement 2 for details).

To test the automated coding procedure and assess pre-service teachers' perception of scaffolding adaptivity, we conducted a pilot study with $N = 16$ pre-service teachers recruited from a seminar and compensated with €10 for voluntary participation. These participants were not involved in either the previous study (Nickl, Sommerhoff, Böheim, et al., 2023) or the main study. They only completed the simulation with adaptive scaffolding. After manually coding their notes, we compared the prompts based on automated coding with those based on manual coding, finding an 80.2% agreement. Participants reported that the prompts supported them where they needed it, but they did not perceive them as personalized. To increase participants' personal relatedness to the prompts, we decided to enrich the adaptive prompts in the main study by adding an explaining sentence ('Based on your prior notes, we found that there is one indicator of mathematical proof skills that you could analyze in greater detail.'), followed by the regular adapted prompt.

4.3. Measures and variables

To measure judgment accuracy, participants evaluated the simulated students' mathematical proof skills using eight items on a 4-point Likert scale, corresponding to the key indicators (three items each for mathematical content and methodological knowledge, and two for problem-solving strategies). An example item is "The student knows what kind of arguments are valid in a proof." These items were intentionally designed to provide a comprehensive assessment of the diverse facets of mathematical proof skills (Chinnappan et al., 2012), encompassing distinct subcategories within mathematical content knowledge, methodological knowledge, and problem-solving strategies (Reiss & Ufer, 2009; see Codreanu, Sommerhoff, Huber, Ufer, & Seidel, 2020). Ratings on the eight items were given for each simulated student at the end of every simulation phase. Judgment accuracy was determined by the number of items where participants' ratings matched those of an expert, allowing for a score range of 0–16 points (up to 8 points per assessed student) in the pre-test, intervention, and post-test simulations, respectively. This procedure of measuring judgment accuracy using the Likert scale was employed and discussed in prior studies (see Codreanu, Sommerhoff, Huber, Ufer, & Seidel, 2020).

4.4. Data analysis

To evaluate our hypotheses regarding the intervention (hypotheses 1a and 2a) and the post-test (hypotheses 1b and 2b), we employed the following analysis scheme. First, we conducted an analysis of covariance (ANCOVA, Type-II), with judgment accuracy as the outcome variable. This analysis used the scaffolding condition (no scaffolding, non-adaptive, and adaptive) as the predictor, and pre-test judgment accuracy as the covariate. We then applied planned contrasts within the ANCOVA model, comparing (i) the judgment accuracy in the no scaffolding condition against both scaffolding conditions (hypotheses 1a and 1b), and (ii) the judgment accuracy in the adaptive condition against the non-adaptive condition (hypotheses 2a and 2b). For each of the two ANCOVA models, we verified the assumptions of homogeneity of variance, independence between condition and pre-test accuracy (as covariate), and homogeneity of regression slopes. None of these assumptions were violated. However, the Shapiro-Wilk test indicated that normality of judgment accuracy in the intervention and the post-test cannot be assumed. With the similar sizes of the conditions, ANCOVA is expected to be robust against violations of normality (Glass, Peckham, & Sanders, 1972). For the sake of completeness, we also conducted non-parametric Kruskal-Wallis test; see electronic supplement 7 for the results. Statistical significance was determined using a significance level of 0.05.

5. Results

Regarding the impact of demographic data on the intervention's effectiveness, no significant interactions with condition and time were found for semester, age, or gender. Two interactions showed isolated cases where judgment accuracy was lower in the non-adaptive condition: for participants from one university (only during the intervention), and for those whose participation wasn't included in a seminar (only during the post-test). This suggests that demographic factors did not systematically influence the intervention's effectiveness.

5.1. Research question

The average judgment accuracy of pre-service teachers in pre-test, intervention, and post-test is shown in Table 1. An ANCOVA analysis for the intervention (hypotheses 1a and 2a) revealed no significant differences in judgment accuracy across all conditions ($F\,(2{,}241) = 0.42$, $p = .658$, $\eta_p^2 < 0.01$). Specifically, the planned contrasts, comparing combined scaffolding (adaptive and non-adaptive) to no scaffolding ($t\,(241) = 0.53$, $p = .600$, $r = 0.03$, hypothesis 1a), and adaptive to non-adaptive scaffolding ($t\,(241) = 0.753$, $p = .452$, $r = 0.05$, hypothesis 2a), were not significant. Likewise, the ANCOVA for post-test judgment accuracy (hypotheses 1b and 2b) showed no significant differences between conditions ($F\,(2{,}241) = 1.20$, $p = .303$, $\eta_p^2 = 0.01$), with neither the contrast scaffolding vs. no scaffolding ($t\,(241) = 0.82$,

Table 1
Mean and standard deviation of participants' judgment accuracy in pre-test, intervention, and post-test.

Condition	*N*	Pre-test	Intervention	Post-test
Control	85	5.46 (2.47)	6.78 (2.49)	6.32 (2.34)
Non-adaptive	81	5.15 (2.30)	6.90 (2.21)	5.74 (2.42)
Adaptive	79	5.27 (2.32)	7.10 (2.30)	6.24 (2.31)

p = .411, $r = 0.05$, hypothesis 1b) nor adaptive vs. non-adaptive scaffolding (t (241) = 1.31, p = .193, $r = 0.08$, hypothesis 2b) reaching significance.

Applying these analyses to the specific indicators of mathematical proof skills yields similar results (see electronic supplement 3 for descriptive results): The indicator-specific ANCOVAs did neither reveal significant differences for the intervention (mathematical content knowledge: F (2,241) = 1.18, p = .310, $\eta_p^2 < 0.01$; methodological knowledge: F (2,241) = 2.32, p = .101, $\eta_p^2 = 0.02$; problem-solving strategies: F (2,241) = 0.20, p = .815, $\eta_p^2 < 0.01$; hypotheses 1a and 2a) nor for the post-test (mathematical content knowledge: F (2,241) = 0.31, p = .736, $\eta_p^2 < 0.01$; methodological knowledge: F (2,241) = 2.06, p = .130, $\eta_p^2 = 0.02$; problem-solving strategies: F (2,241) = 1.61, p = .202, $\eta_p^2 = 0.01$; hypotheses 1b and 2b). Regarding planned contrasts, scaffolding had significant advantages over no scaffolding regarding methodological knowledge in the intervention (t (241) = 2.15, p = .033, $r = 0.14$); all other planned contrasts did not show significant differences (all values of $|t|$ were smaller than 1.61, with $p \geq .109$, $|r| \leq .10$).

5.2. In-depth analyses

Given that our results did not align with the hypotheses regarding our research question, we aimed to uncover potential reasons for this outcome using exploratory analyses. To understand why this adaptation strategy did not significantly affect judgment accuracy, we used the scheme from section 2.4 (see also Fig. 3) for a qualitative, in-depth analysis of adaptive scaffolding in the present study.

To quantitatively describe participants' coverage of the three indicators as the source of adaptation, we calculated *mean coverage* by counting the number of covered indicators in each video (ranging from 0 to 3), summing these counts over all eight videos, and then dividing the sum by the maximum possible count (22). Fig. 4 presents an example of how the adaptation strategy functioned for a participant in the adaptive scaffolding condition. For instance, in the first video of the simulated student Andrew, the participant referred to Andrew's mathematical content knowledge but omitted his problem-solving strategies, so the subsequent prompt focused on the latter. This participant's mean coverage is calculated as $15/22 \approx 0.68$. Similarly, *indicator-wise coverage* can be calculated. For instance, this participant had a coverage of methodological knowledge of $4/7 \approx 0.57$.

5.2.1. Choice of the source of adaptation

First, we focused on the choice of the source of adaptation (here: coverage of the three indicators of mathematical proof skills in participants' notes per video). To check the suitability of this source of adaptation, we followed the scheme (Fig. 3) and assessed if participants' mean coverage relates to their judgment accuracy (see C1 in Fig. 3) and if there is sufficient variability in this coverage (see C2 in Fig. 3) to justify adaptation.

We found a significant but small correlation between mean coverage and judgment accuracy in the pre-test ($r = 0.11$, p = .037, see C1 in Fig. 3). The mean coverage in the pre-test was $M = 0.46$ with substantial variance $SD = 0.19$ (see Fig. 5, see C2 in Fig. 3). Regarding indicator-specific judgment accuracy, covering methodological knowledge in the notes positively influenced judgment accuracy for methodological knowledge ($r = 0.21$, $p < .001$). Similarly, covering problem-solving strategies influenced judgment accuracy for problem-solving strategies ($r = 0.11$, $p = .037$). However, the correlation for mathematical content knowledge was not significant ($r = 0.00$, p = .502). To further assess variability – also with regard to the adaptation strategy, we considered the overlap in the provided scaffolding between the adaptive and non-adaptive scaffolding condition. If this overlap is high, the necessity of adaptivity becomes questionable, and no significant differences between the two conditions are expected, as the treatment is mostly identical. In the adaptive condition, 54% of the provided prompts were equal to those that would have been provided non-adaptively (see electronic supplement 4), indicating a non-neglectable yet not excessive overlap.

5.2.2. Measuring the source of adaptation

The second prerequisite for successful adaptivity implementation involves reliable and valid measurement of the source of adaptation (see M1 in Fig. 3) without disrupting participants' learning (see M2 in Fig. 3). In our study, the coverage of three indicators in participants' notes was measured by automatically coding their notes, ensuring uninterrupted simulation engagement (see M2 in Fig. 3). However, a potential drawback is the reliability of automated natural language processing.

The key question was whether the prompts provided based on automated coding would match those from manual coding. While a pilot of the adaptivity strategy yielded a match of 80.2%, we post-hoc used the data from the current study to re-evaluate the automated coding. To address this, we manually coded notes following our established coding strategy, achieving substantial interrater agreement ($\kappa = 0.71$). We then compared the manually derived prompts with those generated automatically, both following the adaptation logic in Fig. 2. Compared to our pilot study, agreement on provided prompts dropped to 69%, with moderate interrater agreement ($\kappa = 0.56$). The error matrix, comparing prompted indicators based on manual versus automated coding, revealed no systematic discrepancies. For example, out of 79 instances in

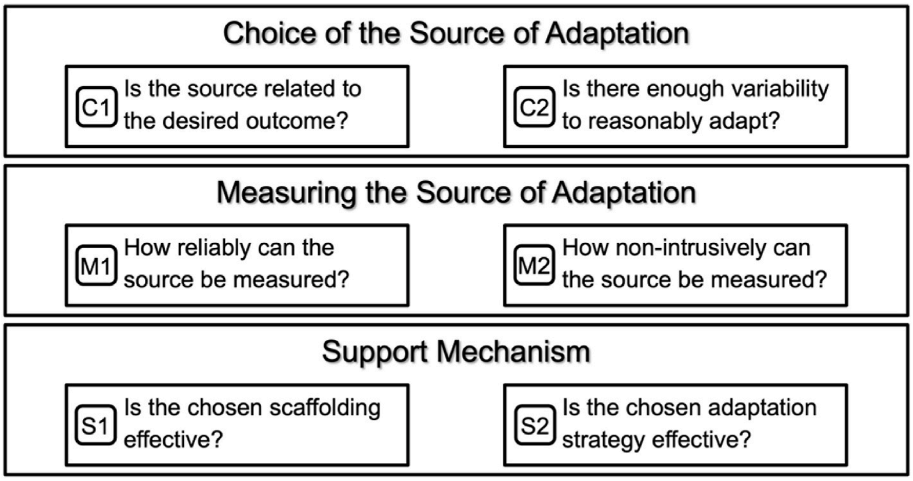
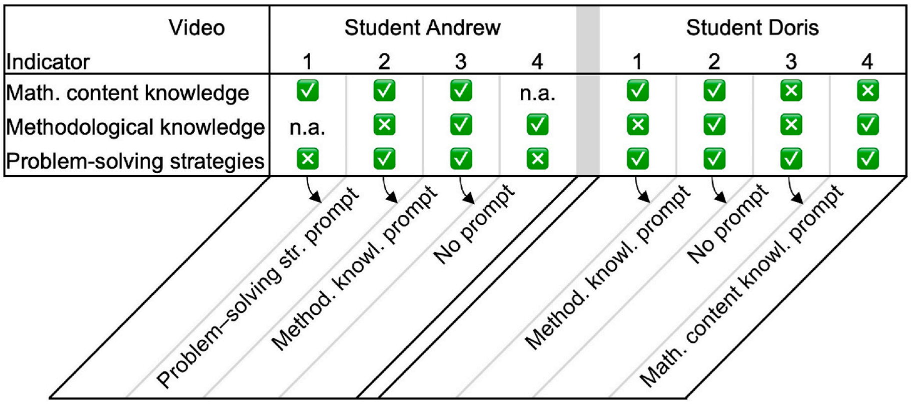

Choice of the Source of Adaptation
C1
Is the source related to the desired outcome?
C2
Is there enough variability to reasonably adapt?
Measuring the Source of Adaptation
M1
How reliably can the source be measured?
M2
How non-intrusively can the source be measured?
Support Mechanism
S1
Is the chosen scaffolding effective?
S2
Is the chosen adaptation strategy effective?

Fig. 3. Scheme for in-depth analysis of adaptive scaffolding.



Video Indicator	Student Andrew 1	2	3	4	Student Doris 1	2	3	4
Math. content knowledge	☑	☑	☑	n.a.	☑	☑	☒	☒
Methodological knowledge	n.a.	☒	☑	☑	☒	☑	☒	☑
Problem-solving strategies	☒	☑	☑	☒	☑	☑	☑	☑

Fig. 4. Coverage of the Three Indicators for Mathematical Proof Skills in Each Video's Notes for a Participant in the Adaptive Scaffolding Condition and Automatically Determined Scaffolding
Note. Indicators lacking cues in the specific video are denoted as 'n.a.'.

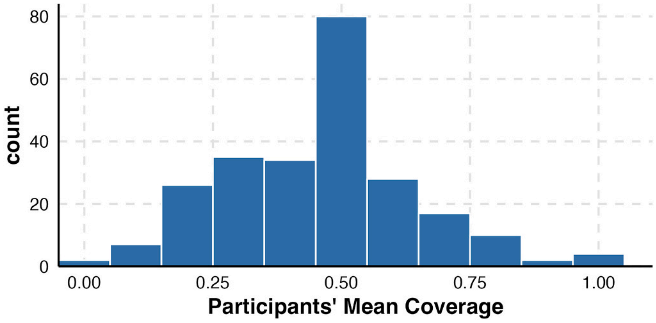


Fig. 5. Histogram of participants' mean coverage.

which prompts on problem-solving strategies were provided, 59 aligned with manual coding, for 3 instances manual coding proposed prompting of mathematical content knowledge, for 7 instances of methodological knowledge, and for 10 instances no prompt (detailed in the electronic supplement 5 for other indicators), indicating that provided prompts fitted participants' notes in most cases.

5.2.3. *Support mechanism*

The third prerequisite for effective adaptation is the efficacy of support mechanisms. This refers to the effectiveness of the chosen (non-adaptive) scaffolding (see S1 in Fig. 3) and the chosen adaptation strategy (see S2 in Fig. 3). The number of provided prompts per indicator can be found in the electronic supplement 4, revealing a focus on methodological knowledge prompts (43.8%) compared to mathematical knowledge prompts (34.7%), problem-solving strategies prompts (9.0%), and no prompt (12.5%).

In our study, the provided prompts aimed to enhance pre-service teachers' note-taking on the previously prompted indicator of mathematical proof skills. Thus, we assessed the scaffolding's effectiveness by comparing the number of sentences about these indicators in the prompted participants' notes to those in the control condition (see S1 in Fig. 3). For example, in Andrew's first video, prompted participants ($N = 160$) wrote on average significantly more on mathematical content knowledge (which was always the prompted indicator for this video) than the 85 participants in the control condition ($W = 4349$, $p < .001$, $|r| = .32$; Wilcoxon rank sum test was used as the Shapiro-Wilk test indicated non-normal distribution). Generally, for each video where a prompt on a specific indicator was given to over ten participants, those prompted wrote more on the targeted indicator than those in the control group (all p's < 0.01, $|r|$'s from 0.21 to 0.47); in cases of $N < 10$, no significant differences were found (see electronic supplement 6 for details).

In our study, prompts were adapted based on participants' coverage of three indicators in their notes, aiming to enhance coverage and consequently as hypothesized, judgment accuracy. Thus, to evaluate the adaptation strategy (see S2 in Fig. 3), we examined differences in the mean coverage during the intervention as a more proximal variable compared to judgment accuracy. Descriptively, the adaptive condition showed the highest coverage, while the control condition had the lowest (see Table 2). A robust Kruskal-Wallis test indicated significant differences in the mean coverage across conditions (H (2) $= 8.25$, $p = .016$), with Bonferroni-corrected Wilcoxon rank sum post hoc tests revealing significant differences between adaptive and control condition ($W = 2504$, $p = .014$, $|r| = .22$). However, differences between non-adaptive versus adaptive ($W = 2839$, $p < .651$, $|r| = .10$) and non-adaptive versus control ($W = 2926$, $p < .281$, $|r| = .13$) were not significant.

Further indicator-wise analyses (see Table 2) showed significant differences in mathematical content knowledge coverage (H (2) $= 14.50$, $p < .001$), with the adaptive condition ($W = 2542$, $p = .019$, $|r| = .21$) and the non-adaptive condition ($W = 2342$, $p < .001$, $|r| = .28$) covering mathematical content knowledge significantly better than the control condition, but no significant difference between adaptive and non-adaptive conditions ($W = 3445$, $p = 1.00$, $|r| = .07$). Coverage of methodological knowledge showed significant differences with both adaptive and non-adaptive conditions significantly covering methodological knowledge better than control condition ($W = 1902$ and $W = 1940$ respectively, both $p < .001$, $|r| = .38$), but without differing significantly from each other ($W = 3083$, $p = 1.00$, $|r| = .03$). Regarding significant differences in covering problem-solving strategies (H (2) $= 12.88$, $p = .002$), the control condition ($W = 4461$, $p = .003$, $|r| = .26$) and the adaptive condition ($W = 2504$, $p = .016$, $|r| = .19$) covered problem-solving strategies significantly better than the non-adaptive condition, with no significant difference between adaptive and control conditions ($W = 3838$, $p = .327$, $|r| = .13$).

Table 2
Mean and standard deviation of coverage measures for the different conditions.

Coverage	Control *M* (*SD*)	Non-adaptive *M* (*SD*)	Adaptive *M* (*SD*)
mean coverage	0.48 (0.16)^{a}	0.53 (0.14)	0.55 (0.15)^{a}
mathematical content knowledge	0.53 (0.22)^{b,c}	0.66 (0.18)^{b}	0.63 (0.20)^{c}
methodological knowledge	0.26 (0.20)^{d,e}	0.43 (0.22)^{d}	0.45 (0.24)^{e}
problem-solving strategies	0.62 (0.27)^{f}	0.50 (0.25)^{f,g}	0.58 (0.23)^{g}

Note. Significant post-hoc tests marked with matching letters.

6. Discussion

While it is theoretically plausible and conceivable that adaptive support should be superior to non-adaptive support, our results did not confirm the anticipated benefits of adaptive scaffolding in enhancing pre-service teachers' assessment skills within our simulation. Neither the impact of scaffolding versus no scaffolding nor the comparison between adaptive and non-adaptive scaffolding showed significant effects regarding judgment accuracy as the central outcome variable. These findings align with the meta-analysis of Belland et al. (2017), which similarly did not corroborate the advantages of adaptive scaffolding suggested by prior theoretical (e.g., Corno, 2008) and empirical work in different contexts (e.g., Ma, Adesope, Nesbit, & Liu, 2014). Belland et al. (2017) did not extensively explore the reasons behind these unexpected results, aside from suggesting the need for further studies to address potential power issues.

The small effects observed in adaptive versus non-adaptive scaffolding required further investigation. To facilitate this, we proposed a scheme (see Fig. 3) to examine adaptive scaffolding. This scheme, encompassing the choice of the source of adaptation, its measurement, and the choice of the support mechanism, offers an approach for predicting the potential success of adaptive support and for understanding variations in its effectiveness. We will discuss each of these components in relation to our study in the subsequent sections.

6.1. *Choice of the source of adaptation*

The significant correlation between coverage and judgment accuracy underpins the validity of coverage as a viable source of adaptation. However, given the small to medium magnitude of the correlations, which also vary by the indicators, and our focus on scaffolding the assessment process (specifically, coverage in participants' notes), it's unlikely that scaffolding-induced improvements in the assessment process would be fully captured by judgment accuracy, particularly in a brief intervention like ours. This is barely surprising as even though judgment accuracy is the standard measure for teachers' assessment skills (Leuders et al., 2022; Urhahne & Wijnia, 2021), it remains a distal measure for capturing pre-service teachers' learning gains regarding the assessment process (Heitzmann et al., 2019; Herppich et al., 2018). Furthermore, effective assessment processes do not always translate into high judgment accuracy (Gabriele et al., 2016).

The overlap of 54% in prompts indicated that on average, half of the prompts in the adaptive condition differed from those in the non-adaptive condition, which likely diminished positive effects of adaptivity on coverage (see S2 in Fig. 3) and judgment accuracy. This overlap relates to 'design-loop adaptivity' (Aleven et al., 2017): the non-adaptive prompts were *designed* in a way that they target the most relevant indicator in the videos, which intentionally coincides with the indicator many pre-service teachers struggle noticing (Sommerhoff et al., 2023). This reduces the additional benefits of real-time adaptivity during the assessment process. This overlap analysis provides a potential measure for determining 'enough variability', a concept not yet consistently defined in current adaptivity research. Future studies should report such quantifiable measures to build substantial evidence for estimating the potential of adaptivity through the variability of data, as suggested by Plass and Pawar (2020). Pre-existing non-adaptive data can be used to anticipate the benefits of adaptivity (e.g., Shute, 1995), but more research is needed to solidify this approach.

Considering alternatives for coverage as a source of adaptation, comprehensive comparisons between different measures are challenging due to a lack of a unified approach to operationalizing the assessment process – theoretically and methodologically: Various frameworks describe partially different aspects of the assessment process (Heitzmann et al., 2019; Herppich et al., 2018; Leuders et al., 2022; van Es & Sherin, 2002), yet they do not specify what constitutes 'good' assessment processes. Methods for capturing the assessment process range from log data analysis (Brandl et al., 2021) to eye-tracking (Kosel, Holzberger, & Seidel, 2021), but lack research into their comparability.

While coverage addresses some of these challenges (Stahnke & Friesen, 2023), an optimal source of adaptation in the context of teachers' assessment skills is lacking. Establishing widely accepted process measures such as coverage, defining the 'optimal' level for these measures (e.g., high coverage), and clarifying the relationship between them (Heitzmann et al., 2019) are promising avenues in future research toward optimizing the choice of the source of adaptation in this context.

6.2. *Measuring the source of adaptation*

In addressing the measurement of the source of adaptation in our study, we focused on reducing intrusiveness. Aligning with the call towards innovative, non-intrusive measures (Shute & Zapata-Rivera, 2012), we emphasized automated natural language processing (Cavalcanti et al., 2021), which can be seen as particularly crucial for evaluating complex skills like assessment skills where problems are multifaceted and typically do not prescribe a unique solution process (Herppich et al., 2018; Leuders et al., 2022). Conversely, using automated coding necessitated a focus on the reliability and validity of these advanced measurement techniques.

Regarding reliability, the agreement levels between manual and automated coding achieved were moderate yet satisfactory compared to those in other contexts ($\kappa = 0.56$ vs. an average $\kappa = 0.40$, and 69% accuracy vs. an average of 68% in cognitive presence, as per Hu, Donald, & Giacaman, 2022). Generative AI, including large language models, holds the potential to further improve the processing of natural language and hence significantly enhance scaffolding strategies in teacher training simulations. When questions of large amounts of training data, data privacy and potential bias can be successfully addressed, large language models can provide enhanced adaptivity and potentially mirror successful incorporation in other educational domains, as highlighted in recent studies on AI-supported essay grading (e.g., Gombert et al., 2024) and personalization of feedback in higher education (e.g., Pfeiffer et al., 2019). However, machine learning approaches, such as those used in automated essay scoring (Ramesh & Sanampudi, 2022), also do not achieve perfect accuracy and can require substantial computational resources.

In this study, perfect accuracy in automated coding may not be crucial. When a mismatched prompt is provided, its impact depends on whether the participant mentioned the indicator in the previous notes. If not (e.g. if a participant only covers mathematical content knowledge in previous notes and should receive a methodological prompt but instead gets a problem-solving strategy prompt), the participant may struggle to assess the prompted more challenging indicator, potentially diminishing the intervention's effectiveness, whereas motivational effects remain positive as the prompt aligns with their needs. If the indicator was previously mentioned (e.g. if a participant covers mathematical content knowledge and methodological knowledge in previous notes and should receive a problem-solving strategy prompt but instead gets a methodological prompt), participants can already assess it, reducing cognitive impact but possibly causing frustration due to the lack of recognition of their progress. However, the neutral formulation of the prompts prevents blame, reducing frustration. Additionally, this scenario is rare since pre-service teachers often focus on only one indicator (Fig. 5), meaning mismatched prompts often still align with their needs.

The validity of the adaptive support in this study is bolstered by its grounding in content-specific theories and empirical evidence. We based the selection of indicators for mathematical proof skills and their assessment through pre-service teachers on prior research (Chinnappan et al., 2012; Codreanu et al., 2021; Nickl, Sommerhoff, Codreanu, et al., 2023). The use of previously validated prompts and the validation of the automated coding in a pilot study ensured our approach's content-specific rigor. This is significant, as many adaptive learning environment methods lack such a detailed content-specific focus

(Gašević et al., 2015; Zawacki-Richter et al., 2019).

6.3. Support mechanism

When focusing on proximal measures in the assessment process such as the number of sentences referring to a specific indicator, scaffolding appeared effective. Apart from cases with $N < 10$, which showed insignificant differences, possibly due to limited power, all comparisons indicated significant effects with small to moderate effect sizes in reaching more processing of cues for the prompted indicators. The adaptation strategy also appeared effective regarding proximal measures of the assessment process, with the adaptive condition outperforming both the control and the non-adaptive condition in covering specific indicators. This effectiveness in our short intervention during the assessment process motivates investigation of longer-term use of prompts, considering their potential interaction with learner prerequisites (Belland, Kim, & Hannafin, 2013), their timing during or after the assessment process (Chernikova, Heitzmann, Fink, et al., 2020), and considering the existence of different prompt types, such as cognitive, metacognitive, or motivational prompts (Bannert, 2009) as well as their varying specificity (Estapa & Amador, 2023).

The effectiveness of the employed scaffolding and adaptation strategy contributes to the overall learning outcomes. Beyond this study's focus on adaptivity, the simulation itself likely facilitates the development of assessment skills (Heitzmann et al., 2019), demonstrating that the control group was comparably strong. Designed to meet pre-service teachers' needs (Codreanu, Sommerhoff, Huber, Ufer, & Seidel, 2020), the simulation's design-loop adaptivity may have additionally lessened the impact of adaptive prompts and contributed to the consistent improvement in judgment accuracy across all conditions from pre-test to post-test. This learning effect aligns with the findings of Chernikova, Heitzmann, Stadler, et al. (2020), which showed that simulation-based learning produced a significant effect size of $g = 0.85$.

Our adaptation strategy, following a recommendation of Aleven et al. (2017), started with simpler tasks of noticing and interpreting mathematical content knowledge and progressed to more complex tasks of noticing and interpreting methodological knowledge or problem-solving strategies. In line with previous research (Nickl, Sommerhoff, Codreanu, et al., 2023), this difficulty rank order in assessing the indicators was also observed in this study (see electronic supplement 3). Additionally, this prompting scheme revealed that most prompts targeted methodological knowledge (43.8%), suggesting that while pre-service teachers are adept at noticing mathematical content knowledge, they struggle more with methodological knowledge. However, past research indicated that knowledgeable learners are more adept at assessing methodological knowledge (Nickl, Sommerhoff, Codreanu, et al., 2023), implying that methodological knowledge and its assessment is learnable (Sporn, 2023). Only a few prompts were provided for problem-solving strategies, suggesting that most learners did not simultaneously process cues for mathematical content and methodological knowledge required to receive such prompts (see Fig. 2 and Aleven et al., 2017). Those who did also often covered problem-solving strategies, as evidenced by the higher frequency of 'no prompt' instances compared to problem-solving prompts, suggesting these participants were relatively proficient.

6.4. Limitations

Although the study was conducted online, which might raise concerns about data quality, we found no indications of compromised data quality. This demonstrates the feasibility of conducting complex research in an online format.

While our study contributes to the understanding of domain-specific assessment processes, particularly regarding mathematical proof skills, its generalizability to other domains may be limited due to the domain-specific nature of assessment skills. Nonetheless, some skills and knowledge (e.g., regarding the role of indicators for the assessment process) may be broadly applicable across various situations (Leuders et al., 2022). If the indicators of the assessed construct are known and their assessment difficulty can be estimated, then transfer of the support mechanism to other assessment contexts is feasible. The sample's heterogeneity, representing various stages of pre-service teachers' studies across different universities, strengthens the ability to generalize the findings on scaffolding adaptivity, but further research is needed for confirmation.

With the heterogeneous sample, we focused on the general effects of scaffolding adaptation, as adaptivity is expected to benefit all learners (Van Schoors et al., 2021). However, differential effects may exist, as prior studies have shown that the effectiveness of prompts is influenced by pre-service teachers' knowledge and motivation (Farrell et al., 2024; Nickl, Sommerhoff, Böheim, et al., 2023; Sommerhoff et al., 2023). A comprehensive moderator analysis involving cognitive and motivational-affective factors is beyond the scope of this article but is a promising direction for future research.

Our focus on notes as a measure of pre-service teachers' assessment process may not have captured their cognitive processes entirely. Typically, pre-service teachers might not document every noticed event and every considered interpretation in their notes. This could have resulted in participants receiving prompts for indicators they considered but had not noted down. The provision of prompts likely mitigated this issue by nudging them to record their thoughts while being carefully worded to minimize frustration if mismatched. Alternative measures such as eye-tracking may offer a more comprehensive view of their cognitive processes through gaze analysis (Kosel et al., 2021). However, eye-tracking also has limitations (e.g., limits regarding the eye-mind hypothesis). Therefore, employing a triangulation of methods to measure the assessment process might provide a more holistic source of adaptation.

Finally, our study mainly focused on judgment accuracy as the primary outcome, operationalized as the alignment of future teachers' solutions with an expert's solution, also aligning with previous research (Urhahne & Wijnia, 2021). Future research could benefit from incorporating additional measures to evaluate various aspects of judgment quality, such as efficiency (Heitzmann et al., 2019), to get a more nuanced understanding of the acquisition of assessment skills through adaptive scaffolding.

7. Conclusion

In this study, we explored the impact of adaptive and non-adaptive scaffolding on enhancing pre-service teachers' assessment skills in mathematical proof skills through a video-based simulation. Utilizing rule-based natural language processing of teachers' notes and informed by domain-specific knowledge, our findings, while not significant in impacting judgment accuracy, align with existing scaffolding research (Belland et al., 2017).

To explore why the intervention did not enhance pre-service teachers' judgment accuracy, we introduced a scheme that provides a systematic approach for an in-depth analysis of adaptive scaffolding. It offers researchers and developers insightful criteria for designing and evaluating the potential of adaptivity of scaffolding. Applying this scheme underscored the necessity of more proximal measures than judgment accuracy for measuring teachers' assessment skills, as well as further intricacies of adaptive scaffolding, such as a notable overlap with 'design-loop adaptivity' (Aleven et al., 2017) or the accuracy of automated coding.

In this regard, our approach to text data analysis marks a significant step in minimizing intrusiveness while maintaining reliability, though it also reveals areas for enhancement, such as achieving higher agreement levels and navigating technical constraints, potentially through machine learning.

Focusing on educational practice in teacher education programs, our study developed and provided an adaptive support tool for teacher training. Moreover, the proposed support mechanism can be transferred to simulations that facilitate pre-service teachers' assessment skills in other assessment contexts.

Overall, our study serves as a valuable addition to the understanding of adaptive, domain-specific real-time scaffolding in teacher education, offering insights and directions for future research and practical application in the field.

Funding

The research was funded by a grant from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) to Tina Seidel, Daniel Sommerhoff, Stefan Ufer, Birgit Neuhaus, and Ralf Schmidmaier (Grant number 1397/11-2, FOR2385).

CRediT authorship contribution statement

Michael Nickl: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniel Sommerhoff:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Anika Radkowitsch:** Writing – review & editing, Methodology, Formal analysis, Data curation, Conceptualization. **Sina A. Huber:** Writing – review & editing, Methodology, Conceptualization. **Elisabeth Bauer:** Writing – review & editing. **Stefan Ufer:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Jan L. Plass:** Writing – review & editing, Conceptualization. **Tina Seidel:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

During the preparation of this work, the author(s) used ChatGPT to improve language and readability of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2024.101994.

References

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522–560). Routledge.

Arnold, K.-H., Gröschner, A., & Hascher, T. (Eds.). (2014). *Schulpraktika in der Lehrerbildung/Pedagogical field experiences in teacher education: Theoretische Grundlagen, Konzeptionen, Prozesse und Effekte/Theoretical foundations, programmes, processes, and effects*. Waxmann. https://elibrary.utb.de/doi/book/10.31244/9783830980575.

Bannert, M. (2009). Promoting self-regulated learning through prompts. *Zeitschrift für Padagogische Psychologie, 23*(2), 139–145. https://doi.org/10.1024/1010-0652.23.2.139

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 505–518). New York: Springer. https://doi.org/10.1007/978-1-4614-3185-5_39.

Belland, B. R., Kim, C., & Hannafin, M. J. (2013). A framework for designing scaffolds that improve motivation and cognition. *Educational Psychologist, 48*(4), 243–270. https://doi.org/10.1080/00461520.2013.838920

Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in stem education: A meta-analysis. *Review of Educational Research, 87*(2), 309–344. https://doi.org/10.3102/0034654316670999

Böttcher, F., & Thiel, F. (2018). Evaluating research-oriented teaching: A new instrument to assess university students' research competences. *Higher Education, 75*(1), 91–110. https://doi.org/10.1007/s10734-017-0128-y

Brandl, L., Richters, C., Radkowitsch, A., Obersteiner, A., & Stadler, M. (2021). Simulation-based learning of complex skills: Predicting performance with theoretically derived process features. *Psychological Test and Assessment Modeling, 63*(4), 542–560.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193–217. https://doi.org/10.1037/h0047470

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., et al. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers & Education: Artificial Intelligence, 2*, Article 100027. https://doi.org/10.1016/j.caeai.2021.100027

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

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Chinnappan, M., Ekanayake, M. B., & Brown, C. (2012). Knowledge use in the construction of geometry proof by sri lankan students. *International Journal of Science and Mathematics Education, 10*(4), 865–887. https://doi.org/10.1007/s10763-011-9298-8

Chou, C.-Y., Lai, K. R., Chao, P.-Y., Tseng, S.-F., & Liao, T.-Y. (2018). A negotiation-based adaptive learning system for regulating help-seeking behaviors. *Computers & Education, 126*, 115–128. https://doi.org/10.1016/j.compedu.2018.07.010

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (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

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (2021). Exploring the process of preservice teachers' diagnostic activities in a video-based simulation. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.626666

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine, 88*(6), 872–883. https://doi.org/10.1097/ACM.0b013e31828ffdcf

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Estapa, A., & Amador, J. M. (2023). A qualitative metasynthesis of video-based prompts and noticing in mathematics education. *Mathematics Education Research Journal, 35*(1), 105–131. https://doi.org/10.1007/s13394-021-00378-7

Farrell, M., Martin, M., Böheim, R., Renkl, A., Rieß, W., Könings, K. D., et al. (2024). Signaling cues and focused prompts for professional vision support: The interplay of instructional design and situational interest in preservice teachers' video analysis. *Instructional Science*. https://doi.org/10.1007/s11251-024-09662-y

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49–60. https://doi.org/10.1016/j.learninstruc.2016.06.008

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review, 16*, 41–67. https://doi.org/10.1016/j.edurev.2015.06.001

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237–288. https://doi.org/10.3102/00346543042003237

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., 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*. https://doi.org/10.1007/s40593-023-00387-6

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055–2100.

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for educational research online, 11*(2), 169–191. https://doi.org/10.25656/01:18004

Heitzmann, N., Seidel, T., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., et al. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda. *Frontline Learning Research, 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*, 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hu, Y., Donald, C., & Giacaman, N. (2022). A revised application of cognitive presence automatic classifiers for MOOCs: A new set of indicators revealed? *International Journal of Educational Technology in Higher Education, 19*(1), 48. https://doi.org/10.1186/s41239-022-00353-7

Kardan, A. A., Aziz, M., & Shahpasand, M. (2015). Adaptive systems: A content analysis on technical side for e-learning environments. *Artificial Intelligence Review, 44*(3), 365–391. https://doi.org/10.1007/s10462-015-9430-1

Kosel, C., Holzberger, D., & Seidel, T. (2021). Identifying expert and novice visual scanpath patterns and their relationship to assessing learning-relevant student characteristics. *Frontiers in Education, 5*, 284. https://doi.org/10.3389/feduc.2020.612175

Kramer, M., Förtsch, C., Stürmer, J., Förtsch, S., Seidel, T., & Neuhaus, B. J. (2020). Measuring biology teachers' professional vision: Development and validation of a video-based assessment tool. *Cogent Education, 7*(1). https://doi.org/10.1080/2331186x.2020.1823155

Kultusministerkonferenz. (2004). *Standards für die Lehrerbildung: Bildungswissenschaften: Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 16.05.2019*.

Leuders, T., Loibl, K., Sommerhoff, D., Herppich, S., & Praetorius, A.-K. (2022). Toward an overarching framework for systematizing research perspectives on diagnostic thinking and practice. *Journal für Mathematik-Didaktik, 43*(1), 13–38. https://doi.org/10.1007/s13138-022-00199-6

Lim, L., Bannert, M., van der Graaf, J., Singh, S., Fan, Y., Surendrannair, S., et al. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior, 139*. https://doi.org/10.1016/j.chb.2022.107547

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123

Meschede, N., Fiebranz, A., Möller, K., & Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: On its relation and differences between pre-service and in-service teachers. *Teaching and Teacher Education, 66*, 158–170. https://doi.org/10.1016/j.tate.2017.04.010

Nakic, J., Granic, A., & Glavinic, V. (2015). Anatomy of student models in adaptive learning systems: A systematic literature review of individual differences from 2001 to 2013. *Journal of Educational Computing Research, 51*(4), 459–489. https://doi.org/10.2190/EC.51.4.e

Nickl, M., Sommerhoff, D., Böheim, R., Ufer, S., & Seidel, T. (2023). Fostering pre-service teachers' assessment skills in a video simulation: Differential effects of a utility value intervention and conceptual knowledge prompts. *Zeitschrift für Pädagogische Psychologie*. https://doi.org/10.1024/1010-0652/a000362

Nickl, M., Sommerhoff, D., Codreanu, E., Ufer, S., & Seidel, T. (2023). The role of teachers' person characteristics for assessing students' proof skills. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, & M. Tabach (Eds.), *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education, 3* pp. 411–418). PME.

Pfeiffer, J., Meyer, C. M., Schulz, C., Kiesewetter, J., Zottmann, J., Sailer, M., et al. (2019). Famulus: Interactive annotation and feedback generation for teaching diagnostic reasoning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP): System demonstrations, Hong Kong, China*.

Pickal, A. J., Engelmann, K., Chinn, C. A., Neuhaus, B. J., Girwidz, R., & Wecker, C. (2023). The diagnosis of scientific reasoning skills: How teachers' professional knowledge predicts their diagnostic accuracy. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1139176

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education, 52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Reiss, K., & Ufer, S. (2009). Was macht mathematisches Arbeiten aus?: Empirische Ergebnisse zum Argumentieren, Begründen und Beweisen [What Makes Mathematical Work? Empirical Results Regarding Proof and Argumentation]. *Jahresbericht der Deutschen Mathematiker Vereinigung, 111*(4), 155–177.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). Macmillan Publishing.

Schons, C., Obersteiner, A., Reinhold, F., Fischer, F., & Reiss, K. (2022). Developing a simulation to foster prospective mathematics teachers' diagnostic competencies: The effects of scaffolding. *Journal für Mathematik-Didaktik*, 1–24. https://doi.org/10.1007/s13138-022-00210-0

Shute, V. J. (1995). Smart: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction, 5*(1), 1–44. https://doi.org/10.1007/bf01101800

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. J. Durlach, & A. M. Lesgold (Eds.), *Adaptive technologies for training and education* (pp. 7–27). Cambridge University Press.

Sommerhoff, D., Codreanu, E., Nickl, M., Ufer, S., & Seidel, T. (2023). Pre-service teachers' learning of diagnostic skills in a video-based simulation: Effects of conceptual vs. interconnecting prompts on judgment accuracy and the diagnostic process. *Learning and Instruction, 101689*. https://doi.org/10.1016/j.learninstruc.2022.101689

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of Teacher Judgments on Student Characteristics and the Construct of Diagnostic Competence]. *Zeitschrift für Padagogische Psychologie, 19*(1/2), 85–95. https://doi.org/10.1024/1010-0652.19.12.85

Sporn, F. (2023). *Mathematisches Beweisverständnis in Sekundarstufe und Hochschule [Doctoral dissertation*. Christian-Albrechts-Universität Kiel. https://macau.uni-kiel.de/receive/macau_mods_00004006.

Stahnke, R., & Friesen, M. (2023). The subject matters for the professional vision of classroom management: An exploratory study with biology and mathematics expert teachers [original research]. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1253459

Stokking, K., Leenders, F., Jong, J.d., & van Tartwijk, J. (2003). From student to teacher: Reducing practice shock and early dropout in the teaching profession. *European Journal of Teacher Education, 26*(3), 329–350. https://doi.org/10.1080/0261976032000128175

Su, J.-M. (2020). A rule-based self-regulated learning assistance scheme to facilitate personalized learning with adaptive scaffoldings: A case study for learning computer software. *Computer Applications in Engineering Education, 28*(3), 536–555. https://doi.org/10.1002/cae.22222

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

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz [The Simulated Classroom: An Experimental Study on Diagnostic Competence]. *Zeitschrift für Padagogische Psychologie, 22*(34), 261–276. https://doi.org/10.1024/1010-0652.22.34.261

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, Article 100374. https://doi.org/10.1016/j.edurev.2020.100374

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education, 10*(4), 571–596. https://www.learntechlib.org/primary/p/9171/.

Van Schoors, R., Elen, J., Raes, A., & Depaepe, F. (2021). An overview of 25 years of research on digital personalised learning in primary and secondary education: A systematic review of conceptual and methodological trends. *British Journal of Educational Technology, 52*(5), 1798–1822. https://doi.org/10.1111/bjet.13148

Vandewaetere, M., & Clarebout, G. (2014). Advanced technologies for personalized learning, instruction, and performance. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 425–437). New York: Springer. https://doi.org/10.1007/978-1-4614-3185-5_34.

Vandewaetere, M., Desmet, P., & Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behavior, 27*(1), 118–130. https://doi.org/10.1016/j.chb.2010.07.038

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Weigand, H.-G., Filler, A., Hölzl, R., Kuntze, S., Ludwig, M., Roth, J., et al. (2014). *Didaktik der Geometrie für die Sekundarstufe I [Geometry Education for Lower Secondary Education]* (2nd ed.). Springer Spektrum. https://doi.org/10.1007/978-3-642-37968-0

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: Types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 35–52. https://doi.org/10.1080/0969594X.2012.725030

Yelland, N., & Masters, J. (2007). Rethinking scaffolding in the information age. *Computers & Education, 48*(3), 362–382. https://doi.org/10.1016/j.compedu.2005.01.010

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Michael Nickl, TUM & IPN, researches effective scaffolding in mathematics teacher education. Daniel Sommerhoff, IPN, investigates mathematical learning and teacher education. Anika Radkowitsch, IPN, explores competences in higher education. Sina Huber applies statistics and machine learning in education. Elisabeth Bauer, TUM, researches adaptive learning support. Stefan Ufer, LMU Munich, investigates mathematical argumentation skills and conceptual understanding. Jan Plass, NYU, investigates learning with digital technologies. Tina Seidel, TUM, researches teacher-student interactions and teacher professional vision.