Generative AI Teaching Simulations as Formative Assessment Tools within Preservice Teacher Preparation

Jamie N. Mikeska[1], Aakanksha Bhatia[2], Shreyashi Halder[1], Tricia Maxwell[1], Beata Beigman Klebanov[1], Benny Longwill[1], Kashish Behl[1], Calli Shekell[3]

[1]ETS Research Institute, [2]ExcelOne, [3]Pennsylvania Western University, Clarion campus

jmikeska,shalder001,tmaxwell,bbeigmanklebanov,blongwill,kbehl@ets.org;
aakankshabhatia01@gmail.com;shekell_c@pennwest.edu

Abstract

This paper examines how a generative AI (GenAI) teaching simulation can be used as a formative assessment tool to gain insight into preservice teachers' (PSTs') instructional abilities. Our team investigated the teaching moves PSTs used to elicit student thinking in a GenAI simulation and their perceptions of the simulation's usefulness.

1 Introduction and Study Aims

Most applications of GenAI in educational contexts during the last year have occurred within K-12 settings, where the primary focus has been on applications that directly support student learning (Chiu, 2025; Mintz et al., 2023). Yet, GenAI also has potential to provide meaningful learning opportunities to teachers to support them in improving their instructional skills, knowledge, and abilities (Lee & Yeo, 2022; Lim et al., 2025; Mikeska & Bhatia, 2025). In this study, our cross-disciplinary team of researchers in teacher learning and educational technology, assessment developers, AI engineers, subject matter experts, and teacher educators collaborated on developing and deploying a GenAI teaching simulation where PSTs could prepare for, engage in, and reflect on their ability to engage in one core teaching practice: elicit and attend to student thinking.

Our team examined how this GenAI teaching simulation could be used as a formative assessment tool to identify the nature of the teaching moves that the PSTs used to elicit and attend to student thinking and the PSTs' perceptions of the simulation's usefulness to support PST teacher learning when integrated within an educator preparation program. By formative assessment, we focus on how the GenAI teaching simulation can be used to gather evidence that can help PSTs understand their instructional strengths and areas for growth and to determine how they could adjust their teaching moves in future instruction (Irons & Elkington, 2021). The main research questions addressed in this study are: (1) What are the teaching moves that elementary PSTs use to elicit and attend to student thinking in a GenAI teaching simulation? and (2) What are PSTs' perceptions of the simulation's usefulness?

2 Background

2.1 Using Digital Teaching Simulations to Support Teacher Learning

While digital teaching simulations can vary in format and structure, most provide PSTs and in-service teachers with opportunities to try out aspects of the teaching within settings of reduced complexity (Dieker et al., 2014; Ersozlu et al., 2021). Digital teaching simulations have been used to support PSTs and in-service teachers in learning how to elicit student thinking, facilitate productive discussions, manage the classroom, and engage with students who are multilingual learners or have special needs (Bondie et al., 2021; Lee et al., 2024; Mikeska et al., 2021). For example, TeachLivE and Mursion use an online simulated classroom that is comprised of up to five student avatars who can interact in real time with the teacher and each other verbally; currently there are multiple simulated classrooms available including an early childhood classroom, upper elementary classroom, middle school classroom, and high school classroom. Other teaching simulations, such as SchoolSims, use an online environment where teachers read through specific scenarios and then are provided

October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

opportunities to make a series of instructional decisions via text-based choices and observe the impact of those decisions.

During the last couple decades, a growing number of research studies have provided empirical evidence illustrating how digital teaching simulations can be integrated productively within educator preparation programs and professional development contexts. Studies have shown that these simulations can be used to improve several outcomes including PSTs' and in-service teachers' ability to engage in core teaching practices, their instructional beliefs, and their content knowledge for teaching (Mikeska et al., 2023; Pecore et al., 2023; Straub et al., 2015). Other studies have suggested that it is important to embed the use of such simulations within learning cycles where teachers have opportunities to prepare for, engage in, and reflect on their simulated teaching experiences, as well as to provide formative feedback to teachers so they can understand and reflect on their instructional strengths and areas for growth (Cohen et al., 2020; McDonald et al., 2013; Mikeska et al., 2021). However, one challenge across this line of research has been the fact that the current simulations require significant human resources to develop and deploy, especially since many of them require a human-in-the-loop to power the student avatars. The recent advances in GenAI offer a potential solution to this challenge – one which we explore in this study by examining the potential of a GenAI teaching simulation as a formative assessment tool with an elementary mathematics methods course.

2.2 Evaluating Teacher Performance

Skilled teaching is critical for positive student outcomes (Blömeke et al., 2016; Fauth et al., 2019). The need for reliable instruments for measuring teacher performance to help them improve has been recognized as a major issue in teacher education research (Correnti et al., 2015). One of the more influential frameworks in this area is the Accountable Talk Theory (Michaels et al., 2008) that provides a protocol for classifying teacher and student contributions to classroom discourse into categories defined by the purpose of each 'move'. For example, teacher talk moves include repeating what the student said and pressing the student for reasoning, while student talk moves include asking for information and relating to what another student said. Talk moves can be reliably identified (Suresh et al., 2022a). More recently, there is work on automating the coding of talk moves and similar constructs to support feedback to teachers (Demszky 2023; Nazaretsky et al., 2023; Suresh et al., 2022b; Tran et al., 2024). Since the protocols are designed to apply across a variety of classroom discussions, eliciting student thinking is only a part of what the teacher does in the bigger picture of facilitating classroom discussions. In this work, we "zoomed in" on the elicitation activity in more detail, since this specific practice is the focus of the simulation. Furthermore, differently from a real classroom, we control the "students" in the simulation by giving them task-specific knowledge profiles that include specific understandings and misunderstandings. As such, we are in a position to evaluate which of the specific points the teacher actually elicited. We therefore used a protocol that combined general categories similar to those in the talk moves literature that pertain to elicitation (e.g., ask questions tied to student actions) and highly content specific categories that focus on unlocking points of understanding or misunderstanding in the simulation (e.g., the student does not understand the commutative property in addition); we call this protocol an "evidence inventory." This two-pronged approach is designed to support feedback both about general tendencies (how often the teacher attends closely to the students' ideas) and about the effectiveness of the elicitation – whether the teacher actually identified the specific pre-designed aspects of the GenAI student's thinking.

3 Study Methodology

3.1 Study Sample

Ten elementary PSTs who were enrolled in an elementary mathematics methods course as part of their educator preparation program at a U.S. university located in the Northeast participated in this study. All PSTs were between 18 to 24 years old and spoke English as their first language. Half of the PSTs had some previous teaching experience via substitute teaching (2 PSTs), as an after school coordinator (1 PST), or as a mentor to elementary students (2 PSTs). None of the PSTs had any previous experience participating in professional learning focused on AI, educational technology, or digital teaching simulations.

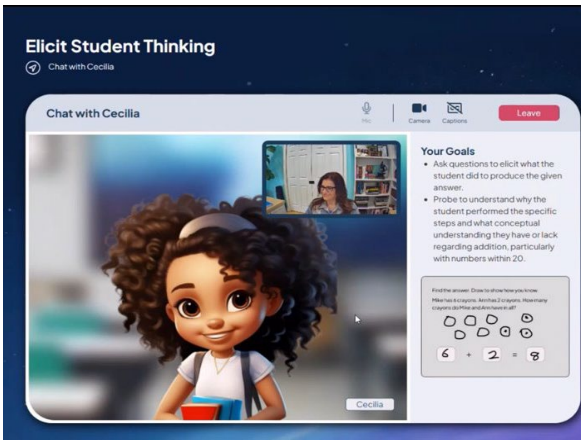
3.2 Data Collection

In this study, the elementary teacher educator integrated the GenAI teaching simulation into their elementary mathematics methods course in Spring 2025 at two different timepoints within a two-week window. At each timepoint, the PSTs had a chance to prepare for, engage in, and then reflect on their GenAI simulated teaching session. Details about the preparation and reflection activities are reported in Mikeska, Beigman Klebanov et al. (2025). Each session used the same GenAI teaching simulation, which we call the Strategies for Adding task.

In this task, PSTs learn that a class of first grade students have been working on learning about strategies for adding numbers within 20 and one student named Cecilia recently solved the following problem: *Mike has 6 crayons. Ann has 8 crayons. How many crayons do they have in all?* The PST's goal in the GenAI simulation is to: (1) ask questions to elicit what the student (Cecilia) did to produce the answer given and (2) probe to understand why the student (Cecilia) performed the particular steps and what conceptual understanding the student has and does not have regarding addition and regarding adding numbers within 20. As part of their preparation, each PST is instructed to review Cecilia's written work (see Figure 1) and prepare by considering ways they could elicit the following: what Cecilia did to produce the answer given, why Cecilia performed the particular steps, and what conceptual understanding Cecilia does and does not have regarding adding numbers within 20, including posing other problems to elicit or confirm Cecilia's understanding.

When ready the PST enters the online environment and begins having a verbal conversation with Cecilia to practice eliciting her thinking about the problem she solved and her understanding in this topic area. Figure 1 shows an image of the Strategies for Adding GenAI simulation interface, as well as shows Cecilia's written work where she drew 8 circles, put dots in three of the circles, and wrote a number sentence underneath the picture (6+2=8).

During the GenAI simulation, all of Cecilia's responses are powered by GenAI. Our team used prompt engineering via GPT-4o on Microsoft's Azure OpenAI service to develop and deploy this GenAI simulation. One of the key resources we leveraged was an already developed human-led simulation task and training protocols, from a previous project, which we then used to develop the initial generation prompt. The initial generation prompt included two parts – instructions and few-shot examples -- to create the response that Cecilia, the GenAI student, would provide during the simulation. Details about the specific prompt used and user testing that our team engaged in to refine the prompt for use within teacher learning contexts can be found in Mikeska, Beigman Klebanov et al. (2025). Previous research indicated that the GenAI student's (Cecilia's) responses in the simulation were: consistently aligned with Cecilia's conceptual understanding and addition problem solving process; age and grade level appropriate; responsive to the teachers' questions and prompts; and coherent across the conversation (Mikeska, Beigman Klebanov et al., 2025).


Elicit Student Thinking

Chat with Cecilia

Chat with Cecilia

Your Goals

- Ask questions to elicit what the student did to produce the given answer.
- Probe to understand why the student performed the specific steps and what conceptual understanding they have or lack regarding addition, particularly with numbers within 20.

Find the answer. Draw to show how you know.
Mike has 6 crayons. Ann has 2 crayons. How many crayons do Mike and Ann have in all?

6 + 2 = 8

Figure 1. Strategies for Adding Online Interface

Chatbot response generation using GPT-4o (v2024-08-06) followed a structured pipeline designed to ensure safety, contextual relevance, and alignment with pedagogical constraints. It began with a request to Microsoft Azure's Chat Completions API, using a system prompt tailored to the GenAI student's profile, few-shot dialogue examples to model interaction style, and the full chat history for context. The API would return a response that has already passed a built-in moderation filter for harmful content. This output then underwent additional validation and transformation steps to reinforce behavioral consistency, ensure educational appropriateness, and reduce the unpredictability of large language model outputs before being presented to the PST.

Primary data sources for this study included written transcripts from each PST's GenAI simulation conversation and survey responses after each session. Each transcript included the utterances from the PST and Cecilia during the conversation (see Appendix A for one example conversation). After each of the two reflection activities, our research team administered an online survey to the PSTs that used both Likert and open-ended questions to gather data about the PSTs' understanding of the GenAI student's thinking and their perceptions of the simulation's authenticity, usability, and usefulness. This study reports on findings from survey questions that asked about the PSTs' perceptions on the usefulness of GenAI teaching simulations within PST learning contexts. Most questions used a Likert scale with Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree as choices for PSTs to select in response to specific statements (e.g., GenAI teaching simulations are a useful tool to support elementary PSTs' learning) while one was an open-ended question about what improvements were needed to the GenAI simulation to best support PST learning.

3.3 Data Analysis

Since each PST engaged in the GenAI simulation at two different timepoints, there were a total of 20 transcripts and survey responses used in the data analysis. To address the first research question, our team used a previously developed evidence inventory rubric to code for the presence or absence of key teaching moves that the PSTs could use in this GenAI simulation to engage in productive aspects of eliciting student thinking. For example, PSTs could use questions or prompts to elicit information about several aspects of Cecilia's problem solving process, including eliciting that Cecilia drew 6 circles and then 2 circles or that Cecilia solved the problem by counting on from 6 (e.g., How did you count to figure out how many crayons they had in all?), and her understanding within this topic area, including that she cannot fluently add the numbers, does not understand the commutative property, and understands what the six, two, and eight represent. These very content-specific categories were coded for Cecilia's turns, namely, where Cecilia's utterance provides evidence that the teacher has successfully elicited this particular element of Cecilia's mathematical thinking. In parallel, PSTs could also use various teaching moves to attend to Cecilia's responses and use them as a basis for further questions, such as asking questions tied to specific things that Cecilia did (e.g., Why did you count on from 6?), and to use follow-up questions or prompts to provide opportunities for Cecilia to explain her reasoning or understanding, such as having Cecilia describe her work and explain aloud (e.g., Why did you only draw dots in three of the circles?). These categories were annotated for the PST's turns and were not tied to the specifics of the mathematical knowledge involved (e.g., "ask Cecilia to explain her reasoning" would be marked the same whether it is about the order of the addends or the use of dots in the circles).

Two raters used the evidence inventory rubric to code for the absence or presence of 18 different teaching moves within the 20 transcripts. If raters noted that specific teaching moves were present in a particular transcript, then they also identified the specific utterances in the transcript that served as evidence of each teaching move. The coding process involved the two raters initially meeting to collectively score one transcript to develop a shared understanding of the 18 teaching moves and the coding process. Then, each rater individually coded the remaining 19 transcripts and then met to reconcile and reach consensus on any individual code applications where they initially disagreed. Overall, the two raters achieved 96.5% exact agreement on the code applications for the presence or absence of these teaching moves across the 19 transcripts and 84.8% agreement for identifying the specific utterances for each teaching move that was identified as present. Finally, we calculated the number and percentage of transcripts that had these teaching moves represented at each timepoint and used the descriptive frequencies to identify the PSTs' strengths and areas for growth within and across timepoints.

To address the second research question, we calculated descriptive frequencies of PSTs' responses to the Likert scale questions about the GenAI simulation's usefulness. Then, we conducted qualitative content analysis (Schreier, 2012) of the PSTs' responses to the open-ended question and calculated descriptive frequencies by codes applied to identify patterns in their responses.

4 Results

4.1 Teaching Moves Used in a GenAI Simulation

Tables 1 and 2 provide the results for the extent to which these PSTs engaged in specific teaching moves, as evidenced by the GenAI student utterances or PST utterances, respectively. These teaching moves were used by the PSTs to elicit the GenAI student's thinking about the process she used and her conceptual understanding about strategies for adding within 20 (Table 1) and to attend to and follow-up on the student's reasoning (Table 2). The results indicate the number and percentage of PSTs (out of 10 PSTs) at each timepoint who exhibited the specific teaching moves in their conversation with the GenAI student. These results indicate several strengths and areas of growth across this group of PSTs.

Table 1. *Teaching Moves to Elicit the GenAI Student's Thinking*

	Teaching Moves (evidenced by the GenAI student utterances)	**Timepoint 1** (n=10 PSTs) n (%)	**Timepoint 2** (n=10 PSTs) n (%)
Focused on the Student's Process	Elicits that the student draws 6 circles and then 2 circles	5 (50%)	6 (60%)
	Elicits that the student draws Mike's crayons first because that is the first number in the problem	0 (0%)	0 (0%)
	Elicits that the student draws Ann's crayons second because that is the second number in the problem	0 (0%)	0 (0%)
	Elicits that the student solves the problem by counting on from 6	5 (50%)	10 (100%)
	Elicits that the student always counts on from one of the numbers in the problem	6 (60%)	10 (100%)
Focused on the Student 's Understanding	Elicits that the student cannot fluently add the numbers	3 (30%)	5 (50%)
	Elicits the student's understanding of the commutative property	1 (10%)	0 (0%)
	Elicits the student's understanding of what the 6 represents	8 (80%)	10 (100%)
	Elicits the student's understanding of what the 2 represents	6 (60%)	9 (90%)
	Elicits the student's understanding of the 8	2 (20%)	1 (10%)
	Elicits the student's understanding that the first addend name summarizes the procedure of counting all of the circles representing that addend	5 (50%)	10 (100%)
	Elicits the student's understanding of the plus symbol	1 (10%)	0 (0%)

First, results suggest that by the second timepoint, all PSTs were able to engage in one or more productive teaching moves to elicit information about the GenAI student's process and conceptual understanding. In particular, the PSTs were most likely to be able to elicit: (a) how Cecilia always counted on from the first addend to solve the addition problem, (b) Cecilia's understanding of what the two addends (six and two) represent, and (c) Cecilia's understanding that the first addend name (six) summarizes the procedure of counting all the circles representing that addend.

For example, one PST asked Cecilia about how she solved the problem and counted; Cecilia replied, "I drew 6 circles for Mike's crayons. Then I drew 2 circles for Ann's crayons. Then I counted 6, 7, 8." Similarly, another PST prompted Cecilia to talk about what the 6 and 2 represented in the number sentence to which Cecilia explained that "Mike's crayons were the first six circles and Ann's were the next 2 circles."

Table 2. *Teaching Moves Used to Attend to and Follow-up on Student's Reasoning*

	Teaching Moves (evidenced by the PST 's utterances)	**Timepoint 1** (n=10 PSTs) n (%)	**Timepoint 2** (n=10 PSTs) n (%)
Focused on the Student's Process	Asks questions tied to specific things that the student did	9 (90%)	10 (100%)
	Attends to and makes use of specific ideas from what the student says	9 (90%)	9 (90%)
Focused on the Student's Understanding	Has the student show work and describe/explain aloud	9 (90%)	10 (100%)
	Poses one or more additional tasks that are clearly useful for the student to solve	1 (10%)	4 (40%)
	Asks questions that lead the student to a particular answer *	3 (30%)	5 (50%)
	Fills in answers for the student (e.g., a contribution that provides information that should have been elicited or probed for) *	1 (10%)	0 (0%)

*These teaching moves do not support the practice of eliciting student thinking.

Second, the results also highlight how these PSTs were quite adept – both at the first and second timepoints – at attending to the GenAI student's idea by asking questions about what Cecilia did and making use of specific ideas that Cecilia shared, as well as using questions to prompt Cecilia to describe and explain aspects of her work. For example, PSTs used various prompts to learn about the steps Cecilia took to solve this addition word problem by asking questions like: "I'd really like to learn too. Can you show me how you're working this problem? What's the first step?"; "Why did you choose that strategy?"; "Can you explain to me why you did the steps you did?"; and "So tell me how did you count on from six?"

Third, the results indicate that there are several areas of growth evident in these PSTs' ability to elicit and attend to student thinking. One of the most striking patterns is that the PSTs were less likely to elicit ideas related to gaps in the GenAI student's conceptual understanding. For example, only one PST (at timepoint 1) was able to successfully elicit that Cecilia did not understand the commutative property (e.g., that 6 + 2 is the same as 2 + 6). Similarly, only 3 PSTs and 5 PSTs at timepoints 1 and 2, respectively, were able to elicit that Cecilia could not do mental math and add numbers fluently in her head; instead, Cecilia always had to draw a picture to represent the addition word problem and then count on from the first addend to solve it.

4.2 Perceptions of the Usefulness of GenAI Simulations

Across both timepoints, most of the PST survey responses to the Likert scale questions indicated that they agreed that GenAI teaching simulations, like the one used in this study, are a useful tool to support elementary PSTs' learning (70% or 14 of 20 PST survey responses across simulation rounds) and can be used to help elementary PSTs better understand student thinking and students' learning needs (85% or 17 of 20 PST survey responses across simulation rounds). There was also strong support that the experience of eliciting student thinking in the simulation closely resembled the work that elementary teachers do to support teaching in real classrooms (75% or 15 of 20 PST survey responses across simulation rounds) and the content addressed in the GenAI simulation was appropriate for elementary PSTs (85% or 17 of 20 PST survey responses across simulation rounds).

In terms of improvements needed to make the GenAI teaching simulation a more effective tool to support PST learning, the qualitative content analysis identified three main ideas. First, in 7 of the 20 survey responses across simulation rounds, PSTs indicated that decreasing the GenAI simulation's latency so that the GenAI student responded more rapidly to the PSTs' questions and prompts would make this tool a more effective one. As one PST noted, "…the only improvement would be the time it took her to respond. The first time, I thought she might not have heard me." Second, in 10 of the 20 survey responses across simulation rounds, PSTs noted that it would be important in GenAI teaching simulations to increase variation in the GenAI student's profile and include additional simulations where students have different conceptual understanding. Finally, in 3 of the 20 survey responses across simulation rounds, PSTs mentioned that the GenAI student's actual responses could be improved to make the simulation more effective, such as by not having Cecilia "repeat herself as much."

5 Conclusion

This study serves as part of broader efforts in the field to determine how GenAI can be used in responsible ways for formative use. Our research context --the use of GenAI to power interactive, online simulations where PSTs can practice and receive formative feedback about their instructional strengths and areas for growth – is one that is currently underexplored, as most research in educational contexts focuses on developing and deploying GenAI tools to support K-12 student learning and outcomes. To ensure that such tools can be used responsibly for formative assessment within teacher learning contexts, a critical first step is ensuring that PSTs' interactions within the GenAI teaching simulations can provide information about PSTs' instructional strengths and areas for growth. It is also important to examine PSTs' perceptions of such tools, as they are more likely to engage with innovative tools if they view them as supportive of their learning.

Findings from this study suggest that GenAI teaching simulations have the potential to be used as formative assessment tools that can be integrated into PST learning contexts. In particular, in this study we developed and deployed a GenAI simulation that provided learning opportunities for PSTs to practice eliciting and attending to student thinking. The study's findings provided empirical evidence of the varied teaching moves these PSTs were able to use successfully to elicit information about the process the GenAI student used to solve the addition word problem and key aspects of her conceptual understanding in this topic area. In addition, the GenAI simulation helped to highlight areas of growth for these PSTs – namely in being able to better elicit gaps in a student's understanding. These findings align with previous research that has indicated teachers struggle to be able to pinpoint challenges that students have and sometimes fail to elicit nuanced information about students' conceptual understanding (Shaughnessy & Boerst, 2018; Sleep & Boerst, 2012).

Results were also promising in terms of the PSTs' mostly positive perceptions about the GenAI simulation's usefulness. Similar results have been reported regarding the use of human-in-the-loop teaching simulations, with results indicating that PSTs and in-service teachers value these simulations to provide content-focused practice spaces where they can improve their instructional capabilities without harming any real students. Ensuring that GenAI simulations provide authentic learning spaces for PSTs that mimic aspects of real classroom interactions is an important step to being able to integrate such tools into PST learning contexts.

Collectively, outcomes from this study suggest that GenAI can be used responsibly to provide a practice-based setting where PSTs can practice eliciting and attending to student thinking, and the outputs of the simulation interaction can be assessed formatively to identify the nature of the teaching moves that the PSTs use – or fail to use – to engage in this instructional practice. This formative information could be used in varied ways to support PST learning, such as incorporating the information into personalized feedback reports for PSTs or having PSTs reflect on the teaching moves they did and did not use to elicit and attend to student thinking after each simulation session. Future research can explore how PSTs make sense of and use this kind of formative information from GenAI teaching simulations to impact their instructional decision-making, can investigate the use of various large language models to power the GenAI student responses, and can examine the use of similar approaches in other content disciplines and topics.

References

Sigrid Blömeke, Rolf Vegar Olsen, and Ute Suhl. 2016. Relation of student achievement to the quality of their teachers and instructional quality. *Teacher Quality Instructional Quality and Student Outcomes*, 2: 21–50.

Rhonda Bondie, Zid Mancenido, and Chris Dede. 2021. Interaction principles for digital puppeteering to promote teacher learning. *Journal of Research on Technology in Education*, *53*(1), 107-123.

Thomas KF Chiu. 2025. Reform, challenges, and future research on AI for K-12 education. *Empowering K-12 Education with AI*. Taylor & Francis.

Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*: 208-231.

Richard Correnti, Mary Kay Stein, Margaret S. Smith, James Scherrer, Margaret McKeown, James Greeno, and Kevin Ashley. 2015. Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. In *Socializing Intelligence through Academic Talk and Dialogue*: 315-334.

Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2023. Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*.

Lisa Dieker, Jacqueline A. Rodriguez, Benjamin Lignugaris/Kraft, Michael C. Hynes, and Charles E. Hughes. 2014. The potential of simulated environments in teacher education: Current and future possibilities. *Teacher Education and Special Education*, 37: 21-33.

Zara Ersozlu, Susan Ledger, and Linda Hobbs. (2021). Virtual simulation in ITE: Technology driven authentic assessment and moderation of practice. In *Authentic Assessment and Evaluation Approaches and Practices in a Digital Era*: 53-68.

Benjamin Fauth, Jasmin Decristan, Anna-Theresia Decker, Gerhard Büttner, Ilonca Hardy, Eckhard Klieme, and Mareike Kunter. 2019. The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and Teacher Education*, 86: 102882.

Alastair Irons and Sam Elkington. 2021. Enhancing earning through formative assessment and feedback (2nd ed.). Routledge.

Dabae Lee and Sheunghyun Yeo. 2022. Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers & Education*, 191: 104646.

Tammy Lee, Carrie Lee, Mark Newton, Paul

Vos, Jennifer Gallagher, Daniel Dickerson, and Camryn Regenthal. 2024. Peer to peer vs. virtual rehearsal simulation rehearsal contexts: Elementary teacher candidates' scientific discourse skills explored. *Journal of Science Teacher Education*, 35: 63-84.

Jieun Lim, Unggi Lee, Junbo Koh, Yeil Jeong, Yunseo Lee, Gyuri Byun, Haewon Jung, Yoonsun Jang, Sanghyeok Lee, and Jewoong Moon. 2025. Development and implementation of a generative artificial intelligence-enhanced simulation to enhance problem-solving skills for pre-service teachers. *Computers & Education,* 232: 105306

Morva McDonald, M., Elham Kazemi, and Sarah Schneider Kavanagh. 2013. Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*: 378-386.

Sarah Michaels, Catherine O'Connor, and Lauren B. Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education,* 27: 283-297.

Jamie N. Mikeska, Beata Beigman Klebanov, Aakanksha Bhatia, Shreyashi Halder, Calli Shekell, Heather Jorgenson, Tricia Maxwell, and Benny Longwill. 2025. Using generative AI digital teaching simulations as practice spaces to support personalized and adaptive learning for preservice teachers in an elementary math methods course. [Manuscript submitted for publication.] Research Institute, ETS.

Jamie N. Mikeska and Aakanksha Bhatia. 2025. Using digital teaching simulations powered by generative artificial intelligence to propel teacher learning. *Journal of the Chartered College of Teaching*. Online.

Jamie N. Mikeska, Dionne Cross Francis, Pamela Lottero-Perdue, Meredith Park Rogers, Calli Shekell, Pavneet Kaur Bharaj, Heather Howell, Adam Maltese, Meredith Thompson, and Justin Reich. 2025. Promoting preservice teachers' facilitation of argumentation in mathematics and science through digital simulations. *Teaching and Teacher Education*, 15: 104858.

Jamie N. Mikeska, Heather Howell, Lisa Dieker, and Mike Hynes. 2021. Understanding the role of simulations in K-12 mathematics and science teacher education: Outcomes from a teacher education simulation conference. Contemporary *Issues in Technology and Teacher Education,* 21: 781-812.

Jamie N Mikeska, Heather Howell, and Devon Kinsey. 2023. Do simulated teaching experiences impact elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics and science? *Journal of Teacher Education*, 74: 422–436.

Joseph Mintz, Wayne Holmes, Leping Liu, and Maria Perez-Ortiz. 2023. Artificial intelligence and K-12 education: Possibilities, pedagogies and risks. *Computers in the Schools,* 40: 325-333.

Tanya Nazaretsky, Jamie N. Mikeska, and Beata Beigman Klebanov. 2023. Empowering teacher learning with ai: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference*, pp. 122-132. 2023.

John Pecore, Corey Nagle, Tadlee Welty, Minkyoung Kim, and Melissa Demetrikopoulos. 2023. Science teacher candidates' questioning and discussion skill performance in a virtual simulation using experiential deliberate practice. *Journal of Science Teacher Education,* 34: 415-435.

Margrit Schreier. 2012. Qualitative content analysis in practice.

Meghan Shaughnessy and Timothy A. Boerst. 2018. Uncovering the skills that preservice teachers bring to teacher education: The practice of eliciting a student's thinking. *Journal of Teacher Education,* 69: 40-55.

Laurie Sleep and Timothy A. Boerst. 2012. Preparing beginning teachers to elicit and interpret students' mathematical thinking. *Teaching and Teacher Education,* 28: 1038-1048.

Carrie Straub, Lisa Dieker, Michael Hynes, and Charlie Hughes. 2015. Using virtual rehearsal in TLE TeachLivE™ mixed reality classroom simulator to determine the effects on the performance of science teachers: A follow-up study. 2015 TeachLivE™ National Research Project: Year 2 Findings.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022a. The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022b. Fine-tuning Transformers with Additional Context to Classify Discursive Moves in Mathematics Classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsey Claire Matsumura. 2024. Multi-dimensional performance analysis of Large Language Models for Classroom Discussion Assessment. *Journal of Educational Data Mining 16*(2), 304-335.

A Appendix

A.1 Example of one PST's transcript

Speaker	Utterance
Teacher	*Hi, Cecilia.*
Cecilia	*Hello!*
Teacher	*Can you walk me through the steps you did to solve the problem?*
Cecilia	*I drew circles for Mike's 6 crayons, then 2 circles for Ann's crayons. Then, 1 counted them: 6, 7, 8.*
Teacher	*Can you tell me why you solved the problem this way?*
Cecilia	*I used circles to show the crayons and counted to find out how many there are altogether.*
Teacher	*How did you know that that would work?*
Cecilia	*I just know it works to count them.*
Teacher	*Why did you decide to use that strategy?*
Cecilia	*Because it helps me see the crayons and count them.*
Teacher	*Can you explain why addition was used here, like needed here?*
Cecilia	*Adding helps me put together Mike's and Ann's crayons.*
Teacher	*So what do you think the problem is really asking?*
Cecilia	*The problem is asking how many crayons Mike and Ann have all together.*
Teacher	*What do you think would happen if you use a different strategy?*
Cecilia	*I don't know how to do that.*
Teacher	*Why did you use counting on here?*
Cecilia	*It makes counting faster. I start at the first number.*
Teacher	*Can you explain a little bit more?*
Cecilia	*I start with Mike's crayons, which are 6, and then count Ann's from there by saying 6, 7, 8.*
Teacher	*So why did you draw five circles with no dots and three with dots?*
Cecilia	*I didn't draw like that. I drew 6 circles and then 2 more. I put dots when 1 counted Ann's crayons after Mike's.*
Teacher	*It looks like in the picture that you drew, it looked like you did five plus three equals eight instead of drawing six plus two.*
Cecilia	*I drew 6 circles for Mike, then 2 for Ann. I just counted on from 6.*
Teacher	*Okay. Good job.*
Cecilia	*Thanks!*

Note. Minor edits have been made to address transcription errors.

Acknowledgement

This work is part of broader efforts in the Praxis program within ETS's K-16 Strategic Business Unit to ensure that GenAI teaching simulations provide productive learning opportunities for educators, including elementary paraeducators, preservice teachers, and in-service teachers. In future work, the Praxis program plans to incorporate a version of the GenAI teaching simulation used in this study into an online learning module designed to support elementary paraeducators in learning how to engage students with subject matter. We appreciate the PSTs who participated in the study and provided us with feedback on the simulation's usefulness.