

Evaluating Educators' Instructional Skills on Making Content Explicit in Instruction in a Generative AI Teaching Simulation

Shreyashi Halder, Jamie N. Mikeska, Heather Jorgenson
ETS

Abstract

This paper examines elementary educators' instructional skills as they practice eliciting student thinking in a generative AI (GenAI) mathematics teaching simulation. Study findings indicate that educators were able to elicit some aspects of a GenAI student's conceptual understanding and misunderstanding. Implications for using GenAI teaching simulations as a practice space and formative assessment tool to build and evaluate educators' instructional skills are addressed.

1 Background

In recent years, there has been a growing effort to improve mathematics teachers' professional competence by enhancing their instructional quality (Blömeke et al., 2015; Orr & Bieda, 2020; Schwartz et al., 2025; Shaughnessy & Boerst, 2018). Strengthening instructional quality does not depend only on cognitive abilities but also on situation-specific facets (Blömeke et al., 2015; Kaiser et al., 2017). Situation-specific facets emphasize teacher attention, interpretation of students' ideas, and instructional decision-making (Santagata et al., 2021); collectively, such facets are referred to as high-leverage teaching practices. High-leverage teaching practices, such as eliciting students' thinking (Orr & Bieda, 2020), facilitating classroom discourse (Chestnutt, 2020), and making content explicit in instruction (Selling, 2016), require thoughtful and skillful tailoring to the specific context and situation (Matsumoto-Royo & Ramírez-Montoya, 2021) to support students' learning in mathematics.

Research has shown that novice teachers struggle more to engage in high-leverage teaching practices than in-service teachers. For example, novice teachers tend to pay attention to the procedural aspects of instruction and struggle to notice critical cues from students' thinking, which can influence their real-time instructional decisions (Arslan et al., 2025; Lisenbee & Tan, 2019). In-service teachers also struggle to use high-leverage practices frequently due to the need to cover required content and meet testing demands (Wilburne et al., 2018). To strengthen their ability to productively use high-leverage teaching practices, teachers need a practice space that replicates authentic experiences and cognitive demands similar to a real classroom, allowing teachers to practice and improve their instructional quality in teaching mathematics (Codreanu et al., 2020; Ferdig et al., 2022; Grossman et al., 2009).

Digital simulations have become increasingly popular as a tool for teachers to enact and reflect on their instructional practices. Digital simulations closely resemble real classrooms but offer reduced complexity (Howell & Mikeska, 2021; Park et al., 2024). However, traditional digital simulations require a human in the loop, called a simulation specialist, who requires training to play student avatars when interacting with the teacher. Requiring a human simulation specialist makes access expensive and limits scalability (Bhowmik et al., 2024; Bondie et al., 2021; Mikeska, Beigman-Klebanov, et al., 2025). Additionally, traditional digital simulations tend to focus more on classroom management (Lindberg & Jönsson, 2023) rather than on specific subjects, such as mathematics (Schwartz et al., 2025).

Recent advances in generative artificial intelligence (GenAI) aim to address these limitations of traditional digital simulations (Tlili et al., 2023; Xu & Zhang, 2023), making them more usable for developing teachers' mathematics instructional practice. GenAI can improve dialogue-based tutoring systems by using large language models (LLMs) to simulate a student during problem-solving interactions (Bhowmik et al., 2024; Macina et al., 2023; Martynova et al., 2025). LLMs can dynamically generate realistic student responses based on students' understanding, misconceptions, and reasoning patterns. Hence, teachers can practice instructional strategies in a controlled yet realistic environment. In particular, GenAI simulations can support mathematics teachers in improving their responsiveness (Son et al., 2024) and creating opportunities to practice elicitation techniques (Mikeska, Bhatia et al., 2025).

These GenAI simulations not only have the potential to support high-leverage teaching practices but also to function as formative assessment tools for evaluating novice teachers' strengths and areas of skill growth (Mikeska, Bhatia et al., 2025; Pilger Suhr et al., 2026). As teacher education programs have transitioned to practice-based learning approaches, there is an increasing need for formative assessments that provide teacher educators with specific insights into novice teachers' mastery of critical instructional skills (Bastian et al., 2018). Formative assessment can continuously support novice teachers' skill development by providing targeted feedback, ultimately preparing them to make informed instructional decisions in real classroom settings. (Pilger Suhr et al., 2026; Villa-Ochoa et al., 2021). Although a few studies have begun to explore GenAI simulations for formative assessment in pre-service teacher education, their application across broader teacher populations remains limited. However, it is necessary to gain in-depth, nuanced insights into teachers' instructional practices and identify specific areas requiring targeted support.

2 Aims

Mathematics teachers require opportunities to practice and master the skills needed to uncover

students' prior ideas. Teachers need opportunities to practice attending to and interpreting students' ideas to make instructional decisions about which ideas are worth exploring and to support students in articulating their thought processes based on their understanding (Orr & Bieda, 2020; Shaughnessy & Boerst, 2018). Halder and colleagues (2026) found that, while interacting with a GenAI-powered student, novice teachers were able to elicit some aspects of a first-grade student's mathematical understanding; however, they struggled to elicit all key understandings and misunderstandings. Furthermore, an analysis of novice mathematics teachers' transcripts revealed that they struggled to ask the GenAI student follow-up questions, which limited their ability to elicit the student's mathematical understanding (Mikeska et al., 2026). These research outcomes are evidence that the development of GenAI simulations has the potential to provide novice teachers with opportunities to try out teaching practices, as well as act as a formative assessment tool to identify teachers' strengths and areas for improvement.

In this study, we examined the use of a GenAI teaching simulation to engage educators in one high-leverage teaching practice when addressing complex mathematical content: making content explicit in instruction. The research questions (RQs) we examined are: 1. To what extent do educators identify the GenAI student's understanding and misunderstanding while making content explicit in instruction? and 2. What variations emerge in educators' ability to make content explicit in instruction during their interactions with a GenAI student?

3 Study Methodology

3.1 Study Context

In this study, participants engaged in a voice-activated mathematics GenAI simulation task entitled *Ordering Fractions*. The GenAI simulation used the GPT-4o model to provide engaging, responsive and advanced conversational capabilities with dynamic interactions for participants. Participants' goal in this simulation task was to explain specific mathematics content to Neil, a third-grade GenAI student. Participants first read the task scenario, which describes Neil's mathematics assignment

and its learning objectives. Neil's assignment is to compare two pairs of fractions: the first pair has a common denominator and different numerators ($\frac{3}{8}$ and $\frac{5}{8}$), while the second pair has a common numerator and different denominators ($\frac{5}{6}$ and $\frac{5}{8}$). Neil correctly compares the first pair of fractions ($\frac{3}{8} < \frac{5}{8}$). He understands that when fractions have the same denominator, the one with the larger numerator is greater. He applies the same whole-number reasoning to compare the second pair of fractions, which is his primary misunderstanding. The educators' goal, therefore, is to use area models to help Neil understand how to compare two pairs of fractions with a common numerator and different denominators.



Area models visually represent fractions by dividing a single shape (e.g., a rectangle) into equal parts based on the denominator and shading parts based on the numerator, making size comparisons easier. In this task, educators use this strategy to help address Neil's misunderstanding and reinforce his learning about how to compare fractions.

Each participant received ten minutes to review Neil's work and the simulation instructions. After completing the preparation phase, the educator begins interacting with Neil (see Figure 1). In this simulation, Neil has a pre- and post-learning phase. In the pre-learning phase, Neil reveals his misunderstanding about using whole number reasoning to compare fractions with the same numerator but different denominators. Then educators must guide Neil in learning how to

correctly compare the fractions using area models. They need to use area models to help Neil understand that as the denominator increases, the size of each individual piece becomes smaller, so for fractions with the same numerator, the one with the larger denominator represents a smaller fraction. Either the participant or Neil can verbalize these relationships to ensure that the learning has taken place. Once this has occurred, Neil enters the post-learning phase where he is able to correctly compare fractions with the same numerators but different denominators.

3.2 Study Sample

Our research team aimed to recruit up to 50 educators, including 20 elementary paraeducators, 20 elementary in-service teachers, and 10 principals who supported paraeducators, to pilot the Ordering Fractions mathematics task. To reach a broad audience, we shared this opportunity with a larger group of educators through a non-profit educational assessment organization's affiliated educator panel. Participants completed an intake and consent form to confirm their eligibility. Inclusion criteria included paraeducators, in-service teachers, or principals with current or previous experience working with paraeducators. We received a response from 645 participants (482 paraeducators, 145 in-service teachers, and 18 principals). From this response pool, we selected 46 paraeducators, 45 in-service teachers, and 16 principals to ensure that there was variability across background characteristics including gender, ethnicity, region, years of teaching experience, and school type. We also ensured that none of them had participated in our previous studies.

A total of 28 participants took part in the study, including 8 elementary paraeducators, 16 elementary in-service teachers, and 4 elementary principals. The participants' ages ranged from 18 to 64. The gender distribution was 82% female and 18% male. In terms of ethnicity, respondents identified as White (57%), Black or African American (21%), Hispanic or Latino (18%), American Indian or Alaskan Native (4%), Asian or Asian American (4%), and Native Hawaiian or Pacific Islander (4%). While 93% of participants had prior experience teaching elementary mathematics, the extent of their experience varied:

15% had less than 1 year of experience, 31% had 1-3 years, 31% had 4-10 years, and 23% had more than 10 years. Regarding familiarity with AI-powered tools in an educational context, participants reported that 42% were not familiar, 15% were a little familiar, 38% were somewhat familiar, and 4% were very familiar.

3.3 Data Collection

Participants completed the simulation and received a feedback report on the various teaching moves they did or did not use during the simulation to make content explicit in instruction. Then they completed a background information questionnaire to provide their educational and professional background and a post-simulation survey (open- and closed-ended items) to report their experience with the GenAI simulation. We collected participants' chat transcripts from the system for further analysis. Two of the open-ended questions analyzed in this study were "explain what the student correctly understands about comparing fractions" and "explain what the student does not understand about comparing fractions." Participants were instructed to use evidence from their conversation with Neil to support their answers.

3.4 Data Analysis

To address RQ1, we used participants' open-ended survey responses detailing what they identified as Neil's understandings and misunderstandings about comparing the fractions.

Table 1. What Neil Does and Does Not Understand

Neil Understands	1. When comparing fractions with the same denominator, the one with the larger numerator is greater.
	2. The numerator tells you how many parts you have.
	3. The denominator tells you how many total parts make up the whole.
	4. How to compare whole numbers (e.g., 101 is greater than 100).
Neil Does Not	1. When the numerator is the same, the fraction with the larger denominator is smaller.
	2. How to compare fractions that have different numerators and different denominators.

Two coders were involved in this process, using a codebook that included Neil's understandings and

misunderstandings, based on the knowledge profile developed for Neil in this task (see Table 1). One coder marked "Yes" if participants identified the understanding demonstrated by Neil and "No" if they did not. A second coder then reviewed this coding. The same coding process was followed for Neil's misunderstandings. Coders convened to resolve any disagreements they noted during the coding. Overall frequencies and percentages for each understanding or misunderstanding aspect across all participants were computed to inform the findings.

Table 2. Aspects and Elements of a Scoring Rubric for Making Content Explicit in Instruction

Aspects	Element Components
1. Introducing and closing the lesson	1. Introduces the particular process, strategy, or technique that is about to be addressed and explains the purpose of using it. 2. Provides or elicits a closing statement that makes explicit important ideas about the targeted process, strategy, or technique.
2. Supporting student learning of the targeted process, strategy, or technique	1. Uses appropriate models, representations, and/or examples to illustrate the targeted process, strategy, or technique. 2. Emphasizes important ideas about the targeted process, strategy, or technique. 3. Uses and explains developmentally-appropriate academic language. 4. Uses a logical sequence and pace, and maintains the instructional focus on the targeted process, strategy, or technique.
3. Checking for student thinking and understanding	1. Elicits student's thinking about the targeted process, strategy, or technique. 2. Uses models, representations, and/or examples to gauge and confirm that student learning about the targeted process, strategy, or technique has been achieved.

To address RQ2, we conducted content analysis of the educators' transcripts (Gläser-Zikuda et al., 2020). We used a holistic scoring rubric (Mikeska, Bhatia et al., 2026; Stickler & Sykes, 2016) to assess educators' overall performance in making content explicit in their conversation with Neil. The scoring rubric includes three aspects, and each has two to four elements (see Table 2).

Three coders were involved in this process; they reviewed the pre-developed scoring rubric, made

modifications as needed to clarify scoring descriptions and levels, and then applied it consistently to score the transcripts. The scoring rubric consisted of three scoring levels (1=novice; 2=beginning; 3=well-prepared). High-performing responses that addressed all required criteria received a score of 3. Responses that partially addressed the required criteria were assigned a score of 2. Responses that minimally addressed the required criteria received a score of 1.

Initially, coders did group scoring using three transcripts to train and develop a shared understanding of the scoring rubric. Two coders continued coding and double-coded four transcripts, achieving an interrater reliability (IRR) of 0.57 (Fleiss's Kappa). All remaining transcripts were double-scored and reconciled by two raters to resolve any disagreements. After scoring the 28 transcripts, we computed the overall frequencies for each scoring level (1, 2, and 3), along with mean and standard deviation across participants per scoring rubric aspect and element.

4 Results

4.1 Identifying the GenAI Student's Understandings and Misunderstandings

The findings from RQ1 (see Table 3) showed that 50% of educators correctly identified the student's understanding that when comparing fractions with the same denominator, the fraction with the larger numerator is greater.

However, only 8%, 15%, and 12% recognized the student's understanding of the numerator, the denominator, and how to compare whole numbers (e.g., 101 is greater than 100), respectively. In identifying the student's misunderstandings, we noted that 35% of educators recognized that Neil thinks that when the numerator is the same, the fraction with the larger denominator is larger. However, none of the educators recognized that Neil does not understand how to compare fractions that have both different numerators and denominators. Overall, the results suggest that while some educators identified some of Neil's key understandings and misunderstandings, many were unable to identify these ideas or gaps.

Table 3. Distribution of Educators Identifying What the GenAI Student Understands and Does Not Understand

Aspects	Para	ISTs	Prin	Total	
	(N=8)	(N=14)	(N=4)	(N=26)	
	n (%)	n (%)	n (%)	n (%)	
Neil Understands	When comparing fractions with the same denominator, the one with the larger numerator is greater.	4 (50%)	8 (57%)	1 (25%)	13 (50%)
	The numerator tells you how many parts you have.	0 (0%)	1 (7%)	1 (25%)	2 (8%)
	The denominator tells you how many total parts make up the whole.	0 (0%)	4 (29%)	0 (0%)	4 (15%)
	How to compare whole numbers	1 (13%)	1 (7%)	1 (25%)	3 (12%)
Neil Does Not Understand	When the numerator is the same, the fraction with the larger denominator is smaller.	2 (25%)	5 (36%)	2 (50%)	9 (35%)
	How to compare fractions that have different numerators and denominators.	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Note: *26 educators completed the post-simulation survey. Para: Paraeducators, ISTs: In-service Teachers, Prin: Principals

4.2 Variations Emerge in Educators' Ability to Make Content Explicit in Instruction

Analysis of RQ2 (see Table 4) showed that aspect one overall, which focused on introducing and closing the lesson, received an average score of 1.5 across educators. This score indicates that educators did not explicitly address the content at the beginning of the lesson by mentioning the key purpose for the lesson and concepts to be addressed, such as "using area models to compare fractions," or closing the lesson by addressing what was learned, such as "for fractions with the same numerator and different denominators, the larger fraction will have the smaller denominator because the parts are larger." However, element-wise, the result indicates that educators were more likely to include a closing statement rather than introducing the lesson.

Table 4. Descriptive Statistics of Aspect and Element Component-wise Chats for Overall Rubric Scores (1 to 3)					
Aspects	EC	Score 1	Score2	Score 3	Mean (SD)
Aspect 1 (N=28)	EC 1	22 (78.6%)	6 (21.4%)	0 (0%)	1.21 (0.42)
	EC 2	12 (42.9%)	5 (17.9%)	11 (39.3%)	1.96 (0.92)
	Overall	15 (53.6%)	12 (42.9%)	1 (3.6%)	1.5 (0.58)
Aspect 2 (N=28)	EC 1	8 (28.6%)	5 (17.9%)	15 (53.6%)	2.25 (0.89)
	EC 2	10 (35.7%)	8 (28.6%)	10 (35.7%)	2 (0.86)
	EC 3	7 (25%)	14 (50%)	7 (25%)	2 (0.72)
	EC 4	11 (39.3%)	7 (25%)	10 (35.7%)	1.96 (0.88)
	Overall	9 (32.1%)	10 (35.7%)	9 (32.1%)	2 (0.82)
Aspect 3 (N=28)	EC 1	19 (67.9%)	5 (17.9%)	4 (14.3%)	1.46 (0.74)
	EC 2	16 (57.1%)	7 (25%)	5 (17.9%)	1.61 (0.79)
	Overall	17 (60.7%)	8 (28.6%)	3 (10.7%)	1.5 (0.69)
<i>Note:</i> EC: Element Component. N=total no. of participants					

Aspect two results indicate that some educators used area models to compare two fractions with the same numerator and different denominators. They explained or guided the student in describing the relationship between numerators and denominators in fractions, using appropriate language, and maintaining a flow of conversation. On the other hand, some educators achieved only partial success, while others may have rarely addressed these key components.

Aspect three, which focused on checking for student thinking and understanding, is similar to aspect one, as both received an overall average score of 1.5. This indicates that these educators rarely elicited students' thinking about the targeted strategy or provided new problems to determine whether the student could apply their new understanding and explain how to compare fractions with the same numerator and different denominators.

5 Discussion and Conclusions

This study demonstrates that dynamic GenAI simulations, such as the one used in this study, offer a potential practice space for mathematics

educators to make content explicit for complex mathematics topics. Educators can use this space to uncover a student's conceptual understandings and misunderstandings and guide the student to address their misunderstandings using effective, grade-appropriate mathematics strategies. The educator can ask follow-up questions or prompts to check the student's new understanding and provide space for the student to articulate their newly acquired insights.

Findings from RQ1 indicated that there were gaps in educators' ability to elicit information during the interaction with the GenAI student, which may prevent them from recognizing the key understandings and misunderstandings that students may have. In particular, these educators often overlooked the key misunderstandings the student had. These findings align with existing research indicating that teachers, particularly novice teachers, often struggle to elicit students' thinking effectively (Arslan et al., 2025; Halder et al., 2026; Mikeska, Bhatia et al., 2025). These gaps indicate that educators might not elicit the full complexity of students' thinking. If educators do not actively engage with students to elicit their thinking, they may not be able to make further instructional moves to help students learn.

Findings from RQ2 suggested that educators frequently struggle to address key components of high-quality lessons, such as introducing the lesson's purpose and summarizing the learning objectives during their instruction. They tend to move directly to address the content rather than sharing the purpose of the lesson, or they abruptly complete the lesson without summarizing the key learning. Moreover, the findings also suggested that educators' follow-up prompts were not as effective in providing students with opportunities to explain their reasoning, which might hinder the elicitation of students' understanding (Mikeska, Bhatia et al., 2026).

This study has several limitations. The sample size was small and the participants only engaged in a one-time simulation, which means we may not be able to generalize the effectiveness of using this simulation for educators to practice making content explicit in their mathematics instruction. A further longitudinal study with a large sample

size could help to explore the broader impact of GenAI simulations. Another limitation of this study was the limited time educators used to address specific lesson content, such as explaining the purpose of the lesson and summarizing it. Due to the simulated nature of the GenAI instructional context, participants may not have considered the importance of making these aspects explicit, like they may do in a real classroom.

Overall, the findings suggest that GenAI simulations may be considered for use as practice spaces and formative assessment tools to evaluate educators' instructional skills in both teacher preparation programs and professional training. However, integrating this simulation alone may not be effective unless teachers receive feedback on their performance. Feedback can be provided through automated feedback reports or through debriefings by teacher educators or coaches to identify teachers' strengths and areas for improvement, ultimately improving the quality of their mathematics instruction.

References

- Arslan, Z., Demirel, D., Çelik, D., & Güler, M. (2025). A study on how novice mathematics teachers respond to high-potential instances of student mathematical thinking. *Thinking Skills and Creativity*, 57. <https://doi.org/10.1016/j.tsc.2025.101859>
- Bastian, K. C., Patterson, K. M., & Pan, Y. (2018). Evaluating teacher preparation programs with teacher evaluation ratings: Implications for program accountability and improvement. *Journal of Teacher Education*, 69(5), 429–447. <https://doi.org/10.1177/0022487117718182>
- Bhowmik, S., West, L., Barrett, A., Zhang, N., Dai, C. P., Sokolij, Z., ... & Ke, F. (2024, September). Evaluation of an LLM-powered student agent for teacher training. In *European Conference on Technology Enhanced Learning* (pp. 68-74). Cham: Springer Nature Switzerland.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13. <https://doi.org/10.1027/2151-2604/a000194>
- Bondie, R., Mancenido, Z., & Dede, C. (2021). Interaction principles for digital puppeteering to promote teacher learning. *Journal of Research on Technology in Education*, 53(1), 107-123. <https://doi.org/10.1080/15391523.2020.1823284>
- Chestnutt, C. (2020, March). Elementary mathematics teacher's knowledge and implementation of high leverage teaching practices. In *Proceedings for the 47th Annual Meeting of the Research Council on Mathematics Learning*.
- 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>
- Ferdig, R. E., Kosko, K. W., & Gandolfi, E. (2022). Using the COVID-19 pandemic to create a vision for XR-based teacher education field experiences. *Journal of Technology and Teacher Education*, 30(2), 239–252. <https://doi.org/10.70725/469822btrvlp>
- Gläser-Zikuda, M., Hagenauer, G., & Stephan, M. (2020). The potential of qualitative content analysis for empirical educational research. In *Forum Qualitative Sozialforschung/ Forum: Qualitative Social Research*, 21(1). <https://doi.org/10.17169/fqs-21.1.3443>
- Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9).
- Halder, S., Mikeska, J. N., Shekell, C., & Jorgenson, H. (2026). Using a Generative AI digital teaching simulation to examine pre-service teachers' skills in eliciting a student's mathematical understanding. *Journal of Mathematics Teacher Education In Texas*, 15(3), 17-25. <https://www.jmtet.net/index.php/Home/article/view/18>
- Howell, H., & Mikeska, J. N. (2021). Approximations of practice as a framework for understanding authenticity in simulations of teaching. *Journal of Research on Technology in Education*, 53(1), 8–20. <https://doi.org/10.1080/15391523.2020.1809033>
- Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M., & Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers—Cognitive versus situated approaches. *Educational Studies in Mathematics*, 94(2), 161-182. <https://doi.org/10.1007/s10649-016-9713-8>
- Lindberg, S., & Jönsson, A. (2023). Pre-service teachers training with avatars: A systematic literature review of human-in-the-loop simulations. *Teacher Education and Special Education. Education Sciences*, 13(8), 817. <https://doi.org/10.3390/educsci13080817>

- Lisenbee, P. S., & Tan, P. (2019). Mentoring novice teachers to advance inclusive mathematics education. *International Journal of Whole Schooling*, 15(1), 1-27. <https://hdl.handle.net/11274/12111>
- Macina, J., Daheim, N., Wang, L., Sinha, T., Kapur, M., Gurevych, I., & Sachan, M. (2023, May). Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2357-2372).
- Martynova, D., Macina, J., Daheim, N., Yalcin, N., Zhang, X., & Sachan, M. (2025, January). Can LLMs effectively simulate human learners? teachers' insights from tutoring LLM students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)* (pp. 100-117).
- Mikeska, J. N., Beigman-Klebanov, B., Bhatia, A., Halder, S., & Suhan, M. (2025, July). Evaluating the use of generative artificial intelligence to support learning opportunities for teachers to practice engaging in key instructional skills. In *International Conference on Artificial Intelligence in Education* (pp. 378-391). Cham: Springer Nature Switzerland.
- Mikeska, J. N., Bhatia, A., Halder, S., Maxwell, T., Klebanov, B. B., Longwill, B., ... & Shekell, C. (2025, October). Generative AI teaching simulations as formative assessment tools within preservice teacher preparation. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers* (pp. 212-220).
- Mikeska, J. N., Bhatia, A., Halder, S., Maxwell, T., Klebanov, B. B., Shekell, C., & Longwill, B. (2026, February). Using a generative artificial intelligence teaching simulation to examine elementary preservice teachers' instructional skills. *Annual Conference of the Association of Mathematics Teacher Educators (AMTE)*
- Orr, S., & Bieda, K. (2020). Preparing to elicit student thinking: Supporting PST questioning in a university teaching experience. In Sacristán, A.I., Cortés-Zavala, J.C. & Ruiz-Arias, P.M. (Eds.). *Mathematics Education Across Cultures: Proceedings of the 42nd Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Mexico* (pp. 1672 – 1676). <https://doi.org/10.51272/pmena.42.2020-264>
- Park Rogers, M., Ogundapo, T., Namakula, E., Lane, K., Cross Francis, D., Bharaj, K., Ataide Pinheiro, P., Maltese, W., Mikeska, A., & Shekell, J. N. C. (2024). Preservice teachers' facilitation of argumentation: Exploring their attention to and perceived complexity of students' thinking. Paper presented at the National Association for Research in Science Teaching (NARST).
- Pilger Suhr, M., Erickson, S., Kelcey, B., Jones, N., Cohen, J., & McLean, L. (2026). Exploring the use of performance tasks as formative assessments of pre-service teacher instructional skills within teacher preparation programs. *Educational Assessment*, 1-23. <https://doi.org/10.1080/10627197.2026.2666090>
- Santagata, R., König, J., Scheiner, T., Nguyen, H., Adleff, A. K., Yang, X., & Kaiser, G. (2021). Mathematics teacher learning to notice: A systematic review of studies of video-based programs. *ZDM – Mathematics Education*, 53(1), 119–134. <https://doi.org/10.1007/s11858-020-01216-z>
- Schwartz, G., Herbst, P., & Brown, A. M. (2025). Harnessing asynchronous digital simulations of problem-based lessons to support mathematics teachers' professional development: A design-based approach. *International Journal of Science and Mathematics Education*, 23(6), 1883-1916. <https://doi.org/10.1007/s10763-024-10514-x>
- Selling, S. K. (2016). Making mathematical practices explicit in urban middle and high school mathematics classrooms. *Journal for Research in Mathematics Education*, 47(5), 505-551. <https://www.jstor.org/stable/10.5951/jresmetheduc.47.5.0505>
- Shaughnessy, M., & Boerst, T. A. (2018). Uncovering the skills that preservice teachers bring to teacher education: The practice of eliciting a student's thinking. *Journal of Teacher Education*, 69(1), 40–55. <https://doi.org/10.1177/0022487117702574>
- Son, T., Yeo, S., & Lee, D. (2024). Exploring elementary preservice teachers' responsive teaching in mathematics through an artificial intelligence-based chatbot. *Teaching and Teacher Education*, 146. <https://doi.org/10.1016/j.tate.2024.104640>
- Stickler, L., & Sykes, G. (2016). Modeling and explaining content: Definition, research support and measurement of the ETS® National Observational Teaching Examination (NOTE) assessment series (Research Memorandum No. RM-16-07). Princeton, NJ: Educational Testing Service.
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., & Huang, R. (2023). ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1), 15. <https://doi.org/10.1186/s40561-023-00237-3>
- Villa-Ochoa, J. A., Sánchez-Cardona, J., & Rendón-Mesa, P. A. (2021). Formative assessment of pre-

service teachers' knowledge on mathematical modeling. *Mathematics*, 9(8), 851.

Wilburne, J., Polly, D., Franz, D., & Wagstaff, D. A. (2018). Mathematics teachers' implementation of high-leverage teaching practices: AQ-sort study. *School Science and Mathematics*, 118(6), 232-243. <https://doi.org/10.1111/ssm.12293>

Appendices

A.1 Partial example of an educator's transcript who scored at level three for all element components of aspect 2

Speaker	Utterance
Teacher	<i>So, I want you to picture two rectangles that are the same size.</i>
Neil	<i>Okay.</i>
Teacher	<i>Take the first rectangle and draw lines up and down to divide it into eight equal parts. Then, in the next rectangle, draw lines up and down to divide it into six equal parts.</i>
Neil	<i>Okay, I'll do that now.</i>
Teacher	<i>So, what do you notice about the rectangle that's divided into eight parts and the rectangle that's divided into six parts?</i>
Neil	<i>The parts in the rectangle with eight are smaller than the parts in the rectangle with six.</i>
Teacher	<i>Exactly. And that's why when the denominators are different, we need a different strategy to be able to compare them. So, you have smaller parts in the rectangle with eight sections. Can you color in five of the parts in the rectangle divided into eight pieces?</i>
Neil	<i>Okay, I'll do that now.</i>
Teacher	<i>Now, color in five of the pieces of the rectangle divided into six pieces.</i>
Neil	<i>Okay, I'll do that now.</i>
Teacher	<i>What do you notice about the parts that you colored in both rectangles?</i>
Neil	<i>The parts I colored in for the rectangle with six pieces are bigger than the parts I colored in for the rectangle with eight pieces.</i>
Teacher	<i>So, if we look back at your work. Which fraction would you say is bigger? Five six or five eight?</i>
Neil	<i>5/6 is bigger.</i>

Note: The teacher's and Neil's statements in the transcript were edited to correct spelling and formatting errors.

A.2 Partial example of an educator’s transcript who scored at level one for all element components of aspect 2

Speaker	Utterance
Teacher	<i>Okay, we can also look at it this way. We can also cross multiply. So if we cross multiply those two numbers. We take eight times three. What is eight times three?</i>
Neil	<i>I don't know how to do that.</i>
Teacher	<i>Okay. Think of your timetables. So what is one times eight?</i>
Neil	<i>One times eight is eight.</i>
Teacher	<i>What is two times eight?</i>
Neil	<i>Two times eight is sixteen.</i>
Teacher	<i>And now what is three times eight?</i>
Neil	<i>Three times eight is twenty-four.</i>
Teacher	<i>Perfect. So I'm going to put a 24 next to the fraction [that] is three eight. Now we're going to keep going. What's eight times four?</i>
<i>Note: The teacher's and Neil's statements in the transcript were edited to correct spelling and formatting errors.</i>	

Acknowledgement

We gratefully acknowledge the support of ETS’s K–16 Strategic Business Unit, whose funding and collaboration made the development and piloting of the generative AI teaching simulation possible. ETS’s K-16 Strategic Business Unit incorporated a version of this simulation into an online learning module designed to support elementary paraeducators in learning how to engage students with subject matter, a high-leverage teaching practice critical for educators’ success in the classroom. We also extend our sincere appreciation to the educators who participated in this study.