Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1460

Practice-Based, Online Modules for Expediting Teacher Skill Development

Julie Cohen
University of Virginia

Nathan Jones
Boston University

Vivian C. Wong
University of Virginia

Alexa Quinn
James Madison University

Steffen Erickson
University of Virginia

Hallie Parten
University of Virginia

Marissa Pilger Suhr
University of Washington

The time available for preservice teacher education is increasingly limited. Teacher preparation programs must find innovative ways to develop teachers' skills within contracted timeframes. One approach is to cover content with online modules. However, most modules teach about skills but do not provide opportunities to practice doing the skills. In this study, we randomly assigned 149 teachers to one of two approaches to online modules: 1) the "business-as-usual" module of readings and reflection questions about skills, and 2) a "practice-based" module including simulated practice sessions with coaching. We provide casual evidence the practice-based module was more effective at developing skills on a range of measures collected in multiple settings, including in student teaching classrooms.

VERSION: April 2026

Suggested citation: Cohen, Julie, Nathan Jones, Vivian C. Wong, Alexa Quinn, Steffen Erickson, Hallie Parten, and Marissa Pilger Suhr. (2026). Practice-Based, Online Modules for Expediting Teacher Skill Development. (EdWorkingPaper: 26-1460). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/vp80-mq53

Practice-Based, Online Modules for Expediting Teacher Skill Development

Julie Cohen

University of Virginia

Nate Jones

Boston University

Vivian C. Wong

University of Virginia

Alexa Quinn

James Madison University

Steffen Erickson

Hallie Parten

University of Virginia

Marissa Pilger Suhr

University of Washington

Acknowledgements: The research presented here was supported by the National Science Foundation (2009939 and 2010298). The opinions expressed are those of the authors and do not represent the views of the National Science Foundation, the Department of Education, or the Institute. We also wish to thank our collaborators Lynsey Gibbons, Robert Q. Berry, Nancy Nelson, Katie Waddell, Ben Kelcey, Lindsey McLean, Christina Taylor, and all the members of the TeachSim lab at UVA. They supported the design and execution of the study, as well as the analyses of the data. Emily Bosch provided exceptional editing support. All errors are those of the authors.

Abstract

The time available for preservice teacher education is increasingly limited. Teacher preparation programs must find innovative ways to develop teachers' skills within contracted timeframes. One approach is to cover content with online modules. However, most modules teach *about* skills but do not provide opportunities to practice *doing* the skills. In this study, we randomly assigned 149 teachers to one of two approaches to online modules: 1) the "business-as-usual" module of readings and reflection questions about skills, and 2) a "practice-based" module including simulated practice sessions with coaching. We provide causal evidence the practice-based module was more effective at developing skills on a range of measures collected in multiple settings, including in student teaching classrooms

The time available for preservice teacher education is increasingly limited. Teacher preparation programs must find innovative ways to develop pre-service teachers' (PSTs') skills effectively and efficiently within contracted timeframes (Grossman et al., 2026). One approach used by both online and university-based programs to cover additional content is integrated asynchronous online modules. However, most of these modules do not address the "problem of enactment" (Kennedy, 1999). They teach *about* classroom skills, but they do not provide opportunities for teachers to practice *doing* the skills. In this study, we randomly assigned 149 individual PSTs within course section blocks to one of two approaches to online modules. One approach was what we consider the "business-as-usual" method for teacher education: reading articles and responding to reflection questions about the focal teaching skill (termed the "reading and reflection module"). The other approach was what we term a "practice-based module" that featured brief simulated practice and coaching sessions about the focal teaching skill (Ball & Forzani, 2009; Grossman et al., 2009). We provide causal evidence that the practice-based module was more effective than the reading and reflection module at developing PSTs' skills on a range of measures collected in multiple settings, including in student teaching classrooms. These results show promise for how teacher education programs could incorporate modularized content in ways that most effectively improves PSTs' skills.

There has been considerable theoretical (e.g., McDonald et al., 2013) and descriptive work (e.g., Kavanagh & Danielson, 2020; Stroupe & Gotwals, 2018) about practice-based teacher education, as well as prominent critiques of the approach (e.g., Kennedy, 2016). However, there has not been any causal research to our knowledge that tests whether practice-based approaches impact teacher skill development in classroom settings (for a systematic review of research about practice-based teacher education, see Cohen et al., 2025a). This is a central contribution of this study.

Our practice-based module featured video examples of the targeted skill, what Grossman and colleagues (2009) call "representations of teaching." The targeted skill was then "decomposed" or analyzed to make its key

features visible. Lastly, the practice-based modules gave teachers an opportunity to try out the target skill, the heart of a practice-based approach, and receive feedback from a live coach (Author, 2026; Grossman et al., 2009). This opportunity to "approximate teaching" was provided through mixed-reality simulations, one kind of approximation that allow PSTs to practice teaching skills in a digitally mediated classroom with student avatars remotely controlled by a live actor (Dieker, 2014). Over the past six years, our research team has demonstrated that the benefits of simulated practice are enhanced when coupled with coaching, but we have not explored whether PSTs transfer those skills outside of the simulation context, including into real classrooms (Cohen et al., 2020, 2023, 2025c). This study extends previous research on mixed-reality simulations by assessing both groups' skills in more standardized settings, using short video-recorded performance tasks (Phelps et al., 2023), as well as in videos of PSTs working in clinical placements (i.e., student teaching) that we scored on a range of observational rubrics.

Both the reading and reflection and practice-based versions of the module focused on metacognitive modeling—or thinking aloud—in elementary mathematics, a high-leverage practice known to support students with disabilities (SWDs) (McLeskey et al., 2017). Supporting SWDs is critical content rarely covered in mathematics methods courses (Jones et al., 2025), making it a good target for modularization. We answer the following research questions:

1. Compared to the business-as-usual teacher education approach of reading and reflection, does a practice-based module focused on metacognitive modeling in mathematics improve PSTs':
    a. Use of the focal skill, metacognitive modeling, in standardized performance tasks?
    b. Use of highly aligned and less aligned measures of teaching skills in clinical placements?

With ongoing teacher shortages and declining enrollment in "traditional" teacher preparation, many teachers are entering the profession without semester or yearlong student teaching placements, let alone placements with an experienced mentor (Kraft & Lyon, 2024; Tooley, 2023). This means many beginning

teachers have limited opportunities to practice and get feedback on their teaching skills (Carver-Thomas et al., 2022; Goldhaber et al., 2020). This study provides causal evidence about an innovative, theory-driven, and potentially scalable method for giving PSTs chances to enact and get feedback on important teaching skills. The fact that the approach caused improvements across a range of skills in real classrooms is hopeful evidence for programs looking to infuse skills-practice without time allocated for student teaching. Our use of a multi-site experimental design – while common in other contexts – is still extremely rare in teacher education research (Hill et al., 2024). In these ways, we demonstrate the potential for not only innovating in teacher education but also innovating in the corresponding research base.

Background and Framework

Teaching is complex work, and it is essential to prepare teachers to do this work well. There have been multiple, distinct approaches to tackling this challenge. One approach has been practice-based teacher education, which infuses a host of pedagogies designed to bolster PSTs' skills. Up until this point, practice-based teacher education has been reliant on highly trained teacher educators, limiting the potential for scale. A distinct approach has been the modularization of content, allowing for asynchronous opportunities to learn about teaching at PSTs' own pace. However, most of the modularized content in teacher education has focused on content delivery and has not infused practice-based methods. Little work to our knowledge has tried to integrate the two approaches by designing practice-based modules that could be flexibly employed in a range of preparatory contexts. Even less work has leveraged causal designs to assess the efficacy of practice-based modules on teacher practice on a range of measures in real classrooms.

Practice-Based Teacher Education

There is considerable debate around how to allocate the scarce instructional time available during pre-service teacher education and what pedagogical methods best support the development of the knowledge, skills, and beliefs undergirding skilled teaching. Some argue case studies are the ideal pedagogy for helping PSTs

grapple with the moral and philosophical dimensions of skilled instruction (e.g., Moreno & Valdez, 2007). Others say PSTs need to read about systemic inequities in K-12 education and reflect on the ways in which those inequities are instantiated in student teaching placements to cultivate asset-based, justice-oriented beliefs about students (Jay & Johnson, 2002; Philip et al., 2019; Zeichner, 2012). Another group of teacher educators have advocated for practice-based pedagogies that afford scaffolded opportunities to: 1) engage with multiple "representations" or examples of teaching skills, 2) analyze or "decompose" those examples, and 3) do or "approximate" the skill with feedback and support from teacher educators (Ball & Forzani, 2009; Lampert et al. 2013; Leko et al., 2023; Schutz et al., 2018). During and after an approximation, teacher educators can pause instructional activities and ask PSTs to analyze aspects of the interactions and reflect on the consequences of their instructional choices and "try again" in ways they could not in real classrooms (Davis et al., 2017; Dutro & Cartun, 2016; Lampert et al., 2013; Richmond et al., 2017).

We still do not know whether the three-part suite of representations, decompositions, and approximations of teaching causes better teaching in real classrooms (Cohen et al., 2025a). Though there has been momentum building for practice-based teacher education, much of the work has been theoretical (Ball & Forzani, 2009; McDonald et al., 2014), or small-scale and descriptive (Kazemi et al., 2016; Schutz et al., 2018). A great deal of the seminal research in the space relies on self-study, perhaps indicating that strong practice-based teacher education is predicated on highly skilled teacher educators in ways that preclude scalability (Kavanagh & Danielson, 2020; Lampert et al., 2013; Stroupe & Gotwals, 2018).

Some causal work has shown that components of practice-based teacher education support positive outcomes for PSTs, though none of the extant research has looked at transfer to real classroom settings. In recent experimental work, Banks and colleagues (2024) found PSTs who had opportunities to decompose the skill of managing student behavior into smaller components (e.g. narrating positive behavior) were more successful at using those management skills in simulated classroom settings. This suggests PSTs benefit from

seeing other teachers engaging in a skill during video examples, but they also need instruction that supports critical analysis around what is happening in those examples and how and why it supports positive student outcomes (Kennedy, 2016). Mancenido and colleagues (2026) randomly assigned college students to different pedagogical approaches to learning how to elicit student thinking. The students who engaged with practice-based approaches performed better on live, standardized "enactments" of similar teaching tasks. Other studies have engaged PSTs in teaching rehearsals and subsequently evaluated their skill development through analysis of lesson plans rather than enacted instruction (Driver et al., 2024; James & Scharmann, 2007). Our lab has conducted a series of randomized control trials looking at the effect of coaching between mixed-reality simulation sessions, compared to self-reflecting between simulations. Across a range of teaching tasks (Cohen et al., 2020, 2023) and teacher preparatory contexts (Cohen et al., 2025c), we find that coaching expedites skill development more than reflection, when we observe PSTs in simulated settings. Taken together, there is suggestive evidence that practice-based approaches can support the development of skilled teaching, but notably, none of the extant literature explores whether practice-based teacher education promotes skill development in live classroom settings, which are invariably more complex. That is a central goal of this study.

Modularizing Content

Given that time is valuable real estate in teacher education, it makes sense there has been a proliferation of modularized content developed for PSTs in recent years. Many of the online programs that dominate the landscape in some states like Texas are composed entirely of asynchronous modules that prospective teachers can move through at their own pace (Kirksey & Gotlieb, 2025). A key advantage of modularized content is the potential for greater flexibility: modules can be embedded within courses, moved between courses, or offered outside of courses.

Interestingly, most of the notable examples of modularized content in teacher education focus on supporting students with disabilities (SWDs). Examples include the dyslexia modules now required in many

states (e.g., South Carolina Department of Education, 2016; Virginia Department of Education, 2016) and the IRIS Center at Vanderbilt's modules, focused on evidence-based practices for supporting SWDs (Sayeski et al., 2015). The IRIS modules provide a range of resources to educators (e.g., real-world scenarios, voiceover commentary from experts, connections to supporting research evidence, built-in assessments) in a self-paced format. These modules are used in most special education teacher education programs, but we know little about whether or how they are used to prepare general education teachers (Smith, Lewis, Montrosse-Moorhead, & Brown, 2014). Moreover, despite the proliferation of the IRIS modules, there is little peer-reviewed research analyzing their efficacy (Sayeski et al., 2015). Kennedy and colleagues have demonstrated the efficacy of content acquisition podcasts (CAPs) that target PST knowledge retention on specific topics, such as vocabulary acquisition (Alves et al., 2018; Kennedy et al., 2016). CAPs, which combine audio and visual input in accordance with Mayer's (2009) instructional design principles, were designed to address limited instructional time in teacher preparation (Kennedy et al., 2011). One additional module-based approach that has emerged in recent years are the toolkits produced by the Institute of Education Science's Regional Educational Laboratories that are aligned to instructional practices highlighted in the What Works Clearinghouse's Practice Guides. Here, too, there is little empirical evidence about whether these modules cause shifts in teachers' skills.

Our modules also focus on differentiating for students with specific needs and scaffolding SWDs, but in addition to learning about differentiation or scaffolding strategies, they provide practice-based opportunities for teachers to learn how to employ these strategies in interactions with students. Our work was grounded in the hypothesis that by designing online, practice-based modules, we could develop not only PST knowledge but also PST skill. Studies of PST skill development have often conflated teacher education content (the knowledge, skills, and dispositions PSTs need to learn) with teacher education pedagogy (how content is taught as part of coursework) (Hill et al., 2024). In their review of special education teacher preparation interventions that target teaching skills, Brownell and colleagues (2020) note existing studies "do little to help us to better

understand *how* to develop the integrated knowledge that expert teachers have for instruction or instructional proficiency in a particular content area" (p. 35). Our research addresses this need, providing some of the first causal evidence about the impact of practice-based modules on teacher skill development.

Modularized Content for Preparing General Education Teachers to Support SWDs in Mathematics

We focus on the preparation of general education teachers to support SWDs in mathematics because there is an urgent need to improve teacher skill in this area and no causal research to our knowledge about how to do so. SWDs, who make up approximately 15% of the K-12 population, often score below their peers on standardized measures of mathematics achievement as early as kindergarten, and these disparities tend to grow over time without targeted support and inclusive instruction (Judge & Watson, 2011; Schulte & Stevens, 2015; Wei et al., 2013). In most cases, SWDs receive their primary mathematics instruction from general education teachers. Unfortunately, these educators largely have been inadequately prepared for the task. Most receive no more than a single course on supporting SWDs (Brownell et al., 2010). General education mathematics methods courses rarely include references to instructional methods known to benefit SWDs, and general education teacher preparation programs almost never have clinical placements focused on working with SWDs (Jones et al., 2026). In an interview study, we found mathematics teacher educators had little knowledge of evidence-based practices for supporting SWDs (Cohen et al., 2025b). Rather than requiring teacher educators to learn new content or dramatically redesign their courses—an unlikely prospect without time and meaningful incentives to do so—we could instead incorporate modularized content into existing courses to provide PSTs more opportunities to develop knowledge and skills for working with SWDs.

We focus on a specific high-leverage teaching practice shown to support SWDs in productively engaging with mathematics word problems: metacognitive modeling (Montague, 1992; Powell & Fuchs, 2018; U.S. Department of Education, 2018). Students with or at risk for mathematics disabilities often struggle to know how to approach mathematics word problems and may over rely on inefficient or less productive

strategies as they work through word problems (Lein et al., 2020). Metacognitive strategies such as self-talk, self-questioning, and self-regulation have consistently been shown to increase student awareness of their own thinking as they solve word problems (Montague & Dietz, 2009; Reid et al., 2013). For example, teachers can teach students to monitor and evaluate their progress toward specific goals and readjust their strategy use as needed by talking themselves through difficult tasks (e.g., asking, "Did I understand what I just read?" or saying, "Next I need to figure out what the important quantities are in the problem."). Teachers can also teach students to monitor and control the emotional strategies they use when they encounter obstacles during problem-solving (e.g., counting to 10 when feeling frustrated). The goal of this practice is not to voice over to students how to solve a problem (i.e., procedural modeling). The goal is to help students slow down and reflect on their own problem-solving process in ways that support persistence and sense of self-efficacy (Powell & Fuchs, 2018). In the strongest examples of metacognitive modeling, the teacher leverages students' own mathematical ideas as a resource for helping them understand how they make sense of and engage with word problems (U.S. Department of Education, 2018).

Our metacognitive modeling modules differ from existing modularized resources about supporting SWDs by focusing on general education teachers. Even though the majority of SWDs spend most of the school day in general education classrooms, general education teachers often graduate from pre-service programs lacking knowledge, skills, and confidence in their ability to meet the needs of SWDs (e.g., CERES Institute for Children and Youth, 2021). Yet there is evidence that interventions during teacher education can improve general education PSTs' attitudes toward inclusion (e.g., Lautenbach & Heyder, 2019), thus establishing a rationale for modularized skill development interventions that target this population. We know of little causal research that tests the effects of introducing modularized content into general education teacher preparation courses, a need our study squarely addresses.

Causal Research in Teacher Education and the Need for Robust Measurement Plans

Experimental research on teacher preparation remains rare, limiting the field's ability to make causal claims about which pedagogies improve preservice teachers' skills (Grossman, 2008; Hill et al., 2024; Mancenido et al., 2024). As a result, many decisions in teacher education continue to rest on tradition, regulation, or local preference rather than a shared empirical base (Brownell et al., 2020; Wilson et al., 2001).

A major barrier to stronger causal evidence is measurement. Clinical assessments are often based on ratings from mentors or supervisors, who may be unreliable assessors of teaching quality (Bartanen & Kwok, 2021). Even trained outside observers may struggle to disentangle PST skill from the influence of mentor teacher expertise (Boguslav & Cohen, 2024). Measurement is also often poorly aligned to intervention targets and timing: preservice teachers are observed only a few times, usually at a remove from their coursework, and programs rarely use standardized assessments before and after specific learning experiences. In addition, many studies rely on project-specific measures that are tightly aligned to intervention content but have limited reliability and validity evidence (Aguilar & Flores, 2022; Mancenido, 2024).

These methodological limitations are important because teacher educators and policymakers need better evidence about whether skills developed during preparation transfer to classroom practice. In prior work, we analyzed the fade-out of skills developed in simulated practice and coaching over the course of a year, but all measurement occurred in simulation sessions, rather than in K-12 classrooms (Cohen et al., 2023). By intentionally using a range of measures across simulated and classroom settings, the present study provides more nuanced information about whether candidates retain skills learned during our instructional modules, whether skill transfer differs based on the pedagogical approach used in the module, and whether there are observable spill overs into the development of additional skills.

Our Study

Experimental Design

During the 2022–2023 academic year, we embedded two-week learning modules into mathematics methods courses at two teacher education programs, in both fall and spring semesters. Within each of the four course sections (or blocks), we randomly assigned individual elementary education PSTs to either a treatment or control condition, yielding a total sample of 149 PSTs across the four blocks. All analyses pool observations across sites, semesters, and blocks.

PSTs in the treatment group (N = 76) learned about our focal teaching skill, metacognitive modeling, through an interactive practice-based module. In this module, PSTs engaged with "representations"—examples of metacognitive modeling—before "approximating" modeling in mixed-reality simulation sessions with coaching support (Grossman et al., 2009). PSTs in the business-as-usual group (N = 73) received the same instructional content, but in a format more typical of teacher education: reading articles, responding to reflection prompts, and receiving feedback on their responses. See Figure 1 for a consort diagram of the study design; see Table 1 for a summary of sample distribution across conditions and sites.

A noted challenge in conducting causal research in teacher education is the design of comparison conditions that adequately address history and maturation threats to validity (Mancenido, 2024). We could have tested our treatment module – focused on supporting students with mathematics difficulties – by comparing outcomes across PSTs who did and did not have any opportunity to learn this content. However, the causal contrast of interest in this study was not content exposure versus no content exposure. Instead, we sought to test whether a practice-based approach to teacher education pedagogy was more helpful for PST skill development than the more modal pedagogy for modularized content, which did not include opportunities to enact targeted skills or receive feedback on them. In this comparison, both groups received the same content, including knowledge of SWDs, mathematics, and metacognitive modeling, but engaged with that content in substantively different ways (Hill et al., 2024).

Description of the Practice-Based Module

We built our practice-based module using Rise360, a web-based e-learning development application from Articulate. It consisted of three related, interactive sub-sections focused on: a) characteristics of SWDs, b)

the complexities of mathematics word problems, and c) the ways in which metacognitive modeling supports student learning in the context of word problems (Powell & Fuchs, 2018). Our approach to online learning design was informed by Moreno and Mayer's (1999) Cognitive Theory of Multimedia Learning. For example, we intentionally chunked information into short segments with more video and less text to minimize cognitive load for PSTs. Interactive features included sorting, matching, and multiple-choice items as low-stakes opportunities to clarify misunderstandings (see Figure 2). Each section ended with a 5-8 item "check for understanding" that required a score of 100% before the PST could proceed to the next section; the interactive content could be revisited as many times as needed to achieve mastery.

The first sub-section focused on building knowledge about students with mathematics difficulties and disabilities. Topics included the medical versus social model of disability (Goering, 2015; Lewis & Lynn, 2018), ableism and intersectionality, and the role of teacher beliefs and expectations in thinking about student "abilities." This was followed by an overview of research on cognition and learning, with a focus on executive functions and learning characteristics of students with mathematics disabilities (Peng et al., 2016; 2018).

The second sub-section was designed to build knowledge about word problems and the challenges students with mathematics difficulties face when navigating these problems (Garcia et al., 2006; Powell et al., 2020). We focused on cognitive load theory, illustrating the high load that word problems present for SWDs (Kennedy & Romig, 2021). This section also included knowledge-building about additive word problem types (Carpenter et al., 1997; Powell & Fuchs, 2018). PSTs learned about strategies elementary students might draw on to "unpack" or make sense of word problems –story context and what the problem is asking students to do– before selecting strategies to actually solve the problem.

The third sub-section focused on metacognitive modeling, making thinking visible to help increase the accessibility of word problems. We first defined metacognitive modeling with examples and nonexamples, including vignettes and classroom videos, what Grossman and colleagues (2009) term "representations of the practice." We created a detailed specification, or "decomposition" (Ibid), of metacognitive modeling that

focused on strategies for thinking aloud about the context of the word problem and the mathematical quantities and relationships within it, detailing the importance of self-questioning, self-monitoring, and self-regulation.

Simulated Teaching Scenarios

All PSTs in the treatment group then participated in two rounds of simulated teaching scenarios with coaching, an approach consistently shown to positively impact PST skill development (Authors, 2020, 2024, 2025). In each of the two rounds, PSTs were provided with an elementary classroom context and mathematical task. Round 1 featured a second-grade task and round 2 featured a fourth-grade task. PSTs were prompted to engage students in making sense of the problem by enacting a metacognitive model in which they think aloud about their thinking related to the problem.

The simulations were digitally mediated, enacted using Mursion software in a Zoom meeting (see Figure 3). Student avatars were remotely controlled by a trained actor, termed an "interactor" (Dieker et al., 2014). The PST could see the student avatars, and the interactor could see the PST, so PSTs could use visual aids as they enacted their metacognitive model. Each simulation round involved a seven-minute practice opportunity, a five-minute coaching session, a second seven-minute practice opportunity, and a final five-minute coaching session. Coaching sessions followed a similar structure. The coach asked an opening question to provide an opportunity for PSTs to self-reflect on their teaching. Coaches then identified an effective element of metacognitive modeling to reinforce (e.g. "You made sense of the overall problem context before discussing the quantities.") Finally, the coach gave a specific next step (e.g., "Don't forget to justify your self-instruction - or explain why these questions are important questions to ask yourself to make sense of the problem.") and explained the rationale behind it, reminding PSTs of the "decompositions" of metacognitive modeling featured in the modules. PSTs were then asked to suggest ways they might incorporate the coach's advice in their next enactment, either during a simulation (after the first round of coaching) or in an elementary classroom (after the second round of coaching).

Description of the Business-as-Usual, Reading and Reflection Module

PSTs assigned to the business-as-usual (i.e., reading and reflection) condition were also assigned to a Rise360 module with three related, interactive sub-sections, focused on the same topics outlined in the description of the treatment module and designed to take the same amount of time (see Figure 4 for an example of what this looked like; see Appendix Table B1 for additional details). To ensure parallel content across conditions, our team carefully reviewed each section of the treatment module, outlined the knowledge and skills introduced, and identified aligned practitioner-oriented articles and response questions. For example, PSTs read a detailed guidance report about metacognition and self-regulated learning that included the recommendation to explicitly teach metacognitive strategies with specific language around how to teach students to plan, monitor, and evaluate their own learning (Quigley et al., 2018). PSTs then responded to open-ended questions such as, "What is modeling and how can it support the teaching of metacognitive regulation?"

Since we were aiming for a business-as-usual approach to teacher education pedagogy, we decided it was important that PSTs in the "reading and reflection" group also receive feedback on their written responses. Our research team generated a feedback bank for each reflection question that included relevant article content, feedback for a strong response, and feedback for a weak response.

For example, for the metacognitive modeling question above, "What is modeling and how can it support the teaching of metacognitive regulation?", an example of a stronger response might be:

> "Modeling is when a teacher or learner explains every step of their thinking in a reflective way when solving problems. It can support the teaching of metacognitive regulation by showing when they stop and think about other ideas for solving a problem, why they did what they did, and maybe even realize they want to use a new strategy when solving. I think the modeling can be reflective and gives a chance to really think about why they did every step and possibly self-correct."
>
> This strong response would receive the feedback "well done," followed by a supplemental explanation:
>
> "The example that folks often use for modeling is of a master craftsperson teaching an apprentice by showing them how they approach the task. Modeling is about making what is implicit explicit so that learners can see examples and build their own skills. The challenge with metacognitive modeling is that what teachers are making explicit is their expert thinking-- and as experts, there are a lot of steps in the problem-solving process that teachers likely fast forward through in their own thinking. Modeling can help teachers promote metacognitive regulation, but it takes a lot of care and thought about what should be included in the model."
>
> Weak responses, on the other hand, might look like:

> "Modeling can help show exactly what the teacher's mental processes are and how to solve the problem,"

The feedback for a weak response would include the same explanation as for a strong response, with the following prompt added at the end:

> "With modeling, we are not trying to create a mathematics classroom where the teacher stands up front and lectures to the students or tells them how to solve a problem. What we are trying to do is to give students an opportunity to see into the black box of the teacher's thinking so they have access to how an expert might approach the same task. We want you to get as much from this activity as possible so please take a moment and look back at Quigley p.16—what is modeling? What isn't modeling?"

Data Collection Protocol

Figure 5 provides an overview of the data collection protocol. At baseline, we collected demographic and background covariates, including PSTs' self-efficacy for teaching students with disabilities (Dawson & Scott, 2013), mathematical knowledge for teaching (Hill et al., 2004), and beliefs about teaching (Clark et al., 2014). We also collected demographic information, including gender, race, age, undergraduate GPA, and PSTs' prior experiences working with students with disabilities and with elementary aged students more broadly.

We also administered an initial performance assessment (Item 1, Figure 5) —consisting of three brief, standardized tasks—to capture PSTs' baseline ability to metacognitively model when unpacking elementary mathematics word problems. These tasks were completed online, on PSTs' personal laptops. For each task, PSTs were provided with a hypothetical classroom context (e.g., "You are working with your second graders on word problems...") and a specific word problem prompt. PSTs were instructed to provide a "think aloud" to model how they would unpack the problem's context, quantities, and mathematical structure. They were also invited to include visual representations as part of their response. See Figure 6 for a screenshot of the task platform. All participants received the same directions for the task, which included a set of key components to include in the model and aligned with the criteria assessed in the performance task rubric. Following completion of the baseline performance tasks, PSTs were randomly assigned to either the control or treatment group.

PSTs in both conditions were given approximately two weeks to complete their assigned module. The first week consisted of ~3 hours working on the asynchronous modules: all three sections for the practice-based group, and sections one and two for the reading and reflection group. The second week consisted of ~1.5 hours:

the third section (focused on the practice) for the "Reading and Reflection" group, and the simulation and coaching sessions for the practice-based group.

Outcome data (classroom recordings and performance task) were collected in a six-week window after PSTs in both groups had completed their respective modules. PSTs were asked to submit a 15–30-minute video of themselves teaching in their clinical placement, which we refer to as the classroom observation. They were instructed to use metacognitive modeling while helping students unpack mathematics word problems, making this a more authentic context for assessing PSTs' use of the focal teaching skill. Before the conclusion of their mathematics methods course, PSTs completed a second set of the standardized performance tasks aligned with the focal instructional practice.

Measures of Teaching Quality

To assess the impact of the two types of modules on PSTs' teaching performance, we analyzed teaching in both standardized performance tasks and clinical placement settings using observation rubrics that varied in their conceptual alignment with the intervention. Below, we describe each measure and associated scoring procedures. Reliability statistics for the measures of teaching quality are presented in Appendix Table A4.

All performance task videos, both pre- and post-engagement with the module, were rated using the researcher-created SimSE rubric, standing for "simulations in special education," the name of our research project. All clinical placement videos were rated on three observation tools: (a) the SimSE rubric, (b) the Quality of Classroom Instruction (QCI; Doabler et al., 2014), and (c) the Mathematical Quality of Instruction (MQI; Hill et al., 2011). The SimSE rubric was the most conceptually aligned to module content. The QCI was moderately aligned. The MQI was the least aligned with the focal skill of metacognitive modeling. This allows us to assess whether effects appeared primarily on highly aligned outcomes or whether they extended to a wider range of classroom-based measures of teaching quality.

SimSE Rubric

Our research team created a rubric focused squarely on the quality of metacognitive modeling, the teaching skill introduced in both the treatment and control modules. Our specification of metacognitive

modeling was developed as part of our team's iterative work to distill the characteristics of a high-quality enactment of the practice. These characteristics are informed by research (e.g., Woodward, 2018), but also by discussions on our team as we reviewed a variety of representations of metacognitive modeling. The rubric contains five items that target key components of a metacognitive model, including stating an objective, quality of unpacking the word problem, quality of self-instruction, quality of self-regulation, and ending the model. Items were rated on a 3-point scale. A rating of 1 represented the lowest score (off-target) while a rating of 3 represented the highest score (on target).

Prior to rating, all raters completed a seven-hour asynchronous training that provided (a) background on the study and the performance tasks, (b) descriptions and written and video examples and non-examples of each rubric item, and (c) opportunities for independent practice rating videos. Following their training, raters were required to complete a reliability checkout and meet at least 0.70 agreement with an expert rating for each of the five rubric items prior to rating independently. Performance task observations were spread across 8 raters and placement video observations were spread across five raters. Raters were randomly assigned to videos to maximize unique combinations of raters and participants, such that raters did not observe the same PST more than once at each assessment timepoint (i.e., pre-test performance task, post-test performance task, and placement video). All raters were blind to participants' condition. Throughout the rating process, raters participated in a biweekly calibration meeting to minimize rater drift, where they were given the opportunity to score and discuss sample performance task videos. Overall SimSE rubric scores were calculated as the mean across items and used as a measure of overall quality of metacognitive modeling. All videos were double scored by two raters and scores from both raters were used in the final analyses.

Quality of Classroom Instruction (QCI)

The QCI measures the quality of teachers' explicit instructional interactions. It captures eight key aspects of explicit instruction: (1) Teacher Modeling, (2) Instructional Pacing, (3) Response Time, (4) Transitions between Activities, (5) Student Engagement, (6) Learning Success, (7) Checks for Understanding, and (8) Encouragement. Each item is rated on a 3-point scale (1 = low quality; 3 = high quality). Total QCI

scores were computed as the mean across all items. A distinct group of trained raters scored clinical placement videos on the QCI in 7.5-minute segments. All videos were double scored, and the total QCI scores from individual lesson segments, scored by each rater, were used in the final analyses.

Mathematical Quality of Instruction (MQI)

The MQI measures the quality and nature of mathematics instruction and was used to score teachers on five domains: (1) Common Core-Aligned Student Practices, (2) Working with Students and Mathematics, (3) Richness of Mathematics, (4) Errors and Imprecision, and (5) Classroom Work is Connected to Mathematics. Within those dimensions, segments of the video are rated on a 4-point scale (0 = Not Present, 1 = Low, 2 = Mid, 3 = High) for several items that support each dimension. Additionally, whole videos are rated on 10 dimensions (e.g., "Lesson time is used efficiently", "Lesson is mathematically designed") on a 5-point scale (1 = Not at all true of this lesson, 3 = Default score, 5 = Very true of this lesson). The SimSE project recruited two highly qualified MQI raters who were already trained and certified to use the MQI rubric and who continued to rate and provide instructional coaching related to the dimensions outside of this project. All videos were double-scored, and both raters were assigned all videos in random order. Total MQI scores were calculated as the mean across domains. Total MQI scores from individual lesson segments, scored by each rater, were used in the final analyses. The raters met biweekly to recalibrate by watching a lesson from the MQI library and reviewing their codes and how they aligned with the codes from the MQI rating library.

We hypothesized the practice-based module would be more effective than the reading and reflection module at improving observable teaching skills because it allowed PSTs to practice those skills in a low-stakes simulated setting and receive feedback before working with students in classrooms. We anticipated the largest effects on outcomes more proximal to the intervention, standardized performance tasks we describe in detail below, and smaller effects in settings less proximal to the intervention, student teaching placements. We also hypothesized we would observe the largest treatment effects on rubrics most conceptually aligned with the content covered in the modules, with smaller effects on rubrics less tightly aligned to module content. This

study illustrates the value of a prospective measurement plan that includes outcomes collected across multiple settings and with varying degrees of conceptual alignment to the intervention.

Analytic Strategy

Before estimating treatment effects, we implemented a series of validity checks to ensure the integrity of the randomized controlled trial. These included assessing baseline balance across treatment arms on pre-treatment covariates, examining attrition and missingness patterns, and documenting any cases of treatment non-compliance. As described in the Results section, we find no statistically significant differences between groups on baseline characteristics, minimal attrition, and only one case of non-compliance. These findings support the internal validity of the design and justify the use of intent-to-treat (ITT) analyses.

Our analytic strategy was designed to capture the effect of our practice-based learning module—the intervention of interest— on PSTs' skill development, compared to our reading and reflection module, the business-as-usual approach to teacher education pedagogy, on the range of outcome measures described above. We employed a blocked randomized design, with random assignment of individual PSTs conducted within mathematics methods course sections.

Estimating treatment effects from observational ratings can be challenging due to multiple sources of error that may bias or obscure the true effect. Bias may arise if PSTs in one condition are disproportionately assigned more favorable raters or encounter more optimal lesson conditions. Imprecision can occur if the measurement tools capture noise due to contextual variation (e.g. classroom dynamics, video quality) rather than true skill.

To mitigate these threats, we used a multi-faceted measurement design (Brennan,1992) in which each PST was rated by multiple trained observers across multiple tasks (for performance assessments) or lesson segments (for classroom observations). Raters were randomly assigned to participants such that each PST was scored by a unique pair of raters, and rater assignments were balanced across treatment arms. This design increases both the precision and the validity of the estimated treatment effects.

Despite these efforts, deviations from the planned measurement design occurred. Some PSTs did not complete all assigned performance tasks, and some raters scored more or fewer videos due to logistical constraints. Additionally, we did not control for lesson length in the classroom observation protocol, and PSTs in the treatment group delivered slightly longer lessons on average – resulting in more observed segments. If lesson length is correlated with the classroom score outcomes, the treatment-control imbalance could bias treatment effect estimates in either direction.

To account for these sources of error variation, we estimate treatment effects using a long-format dataset organized at the participant-by-lesson-segment (or task)-by-rater-level, with multiple outcome records per participant. This specification allowed us to include fixed effects for raters and lesson segments (or task) in the regression models. Including these controls help address imbalance in the number of observed segments and potential heterogeneity in scoring across raters and tasks.

More formally, we estimated the following least squares regression model:

$$Y_{\mathrm{ijb}}=\beta_0+\beta_1 \mathrm{Treatment}_{\mathrm{i}}+\gamma_j+\delta_k+\lambda_b+\mathrm{X}_{\mathrm{i}}'\theta+\epsilon_{ijb}$$

Where Y_{ijb} is the score for PST i, rated by rater j, in lesson segment or task k, within randomization block b. The model includes fixed effects for raters (γ_j), segments or tasks (δ_k)[1], and blocks (λ_b).

To improve the precision, we also include a vector of baseline covariates Xi, including: pre-test performance task scores (SimSE rubric), Mathematical Knowledge for Teaching (MKT), teaching self-efficacy, NEO personality traits, undergraduate GPA, prior teaching experience, teaching beliefs, gender, and race. Reliability statistics for the survey measures used as covariates are presented in Appendix Table A3. Standard errors are clustered at the PST level to account for repeated observations. As a robustness check, we run additional models without baseline covariates or rater and segment (task) fixed effects to assess potential treatment-control imbalances and test the sensitivity of the results to model specification.

Findings

Sample Description

[1] We additionally ran models where we averaged outcomes over the full video and they were consistent with the main findings.

Table 1 summarizes sample characteristics across the four course sections in which the intervention was implemented. The average PST was 20.8 years old, with a college GPA of 3.59. The sample was predominantly White (78%) and female (99%). On average, PSTs entered the study with high teaching self-efficacy and roughly two years of prior experience working with young people (i.e., tutoring, working as a camp counselor etc.), but low baseline scores on the SimSE rubric and limited mathematical knowledge for teaching. While some variation is visible across blocks (e.g., Site 1 Fall PSTs were older and reported lower self-efficacy), all analyses pool observations and include block fixed effects to account for these differences.

Validity of Experimental Design

We conducted several diagnostic checks to assess the integrity of the randomized design. Table 2 shows no statistically significant differences between treatment and control groups across any measured baseline covariates. The joint test of covariate balance also fails to reject the null ($F(9, 144) = 0.45$, $p = 0.902$), providing strong evidence of successful random assignment. As shown in Figure 1, our consort diagram, of the 149 PSTs randomized, 146 completed both pre- and post-intervention performance tasks, and 140 were observed in their clinical placements. In Appendix Table A2, we present balance tables for the baseline, performance task, and clinical placement analytic samples. One PST assigned to treatment withdrew before engaging in study activities, representing a single case of treatment non-compliance. Accordingly, we present intent-to-treat (ITT) estimates in the main analyses.

Experimental ITT Effects

We find statistically significant and substantively meaningful treatment effects across all four outcome measures—ranging from those most closely aligned with the module content (performance tasks scored using the SimSE rubric) to more distal and less tightly aligned outcomes (classroom observations scored with the QCI and MQI rubrics). This pattern aligns with theoretical expectations: the largest effects occur on measures conceptually proximal to the intervention, with smaller effects observed on less aligned outcomes (see Figure 7). This descending pattern of effects provides support not only for the internal validity of the study, but also for

the construct validity of the outcome measures and the theory of proximal transfer—where we expect the strongest gains on directly taught skills and more modest effects on near and far-transfer outcomes.

The results from our main specification are presented in Table 3. We also present TOT effects in Appendix Table A1. The largest effect is observed on PSTs' metacognitive modeling during performance tasks as rated on the SimSE rubric ($b = 0.961$, $SE = 0.11$, $p < .001$), representing nearly a full standard deviation improvement relative to the control group mean (Table 4, $M = 1.31$, $SD = 0.22$). This suggests that the intervention had a large impact on PSTs' ability to produce structured metacognitive models in a standardized assessment setting.

The next largest effect is observed on PSTs' classroom metacognitive modeling, as rated on the same SimSE rubric ($b = 0.769$, $SE = 0.12$, $p < .001$). Although smaller than the effect on the performance tasks, it still represents a substantial gain, especially given the additional complexity of classroom instruction. These findings indicate that the practice-based module better supported PSTs in both learning and enacting the focal teaching skill across distinct instructional contexts.

Smaller but still statistically significant effects were found on two broader measures of instructional quality. The average effect on PSTs' QCI scores—a measure of explicit instruction—was 0.362 SD ($SE = 0.12$, $p = .003$). The QCI includes items related to modeling, pacing, and student engagement, and was developed by researchers focused on supporting students with disabilities in inclusive math classrooms (Doabler et al., 2014). Its conceptual alignment with the focal practice of metacognitive modeling may explain the moderately sized effect.

The smallest treatment effect was observed on MQI scores ($b = 0.261$, $SE = 0.14$, $p = 0.064$), which falls slightly short of conventional thresholds for statistical significance. However, the estimate is consistent with the overall pattern and may reflect broader instructional improvements, even in areas not explicitly targeted by the practice-based module. The MQI emphasizes the clarity and accuracy of mathematical explanations and the cognitive demand of tasks—constructs not directly targeted in our modules. This modest effect suggests the practice-based module may have supported improvements in general instructional quality, even beyond the

focal skill. Given the relatively small sample size and consistent pattern of findings, we interpret this result as substantively meaningful.

The results are largely robust to the choice of model specification, including whether baseline covariates or rater and segment (or task) fixed effects are included in the models. The results from different model specifications are presented in Appendix Tables A5 and A6. Without covariates, the effects are attenuated slightly but the pattern of effects, and their statistical significance remain consistent. The one exception is that in models without rater fixed effects, the treatment effect on the Classroom QCI outcome becomes insignificant. This occurs because some QCI raters exhibited systematic scoring biases, and raters with lower scoring tendencies were disproportionately assigned to the practice-based module group, whereas raters with higher scoring tendencies were disproportionately assigned to the reading and reflection module group. The rater fixed effects control for these imbalances.

The results are also robust to differences in attrition across the two groups. To assess the impact of attrition on the treatment effect estimates, we calculated Lee bounds with 95% confidence intervals using the Stata "leebounds" command. Because the command does not allow users to include continuous covariates, we estimated Lee bounds using the regression specification that excluded covariates but controlled for rater, segment (or task), and randomization block fixed effects. Lee bounds are presented in Appendix Table A7 and should be compared to Appendix Table A5. The results are largely consistent with the main findings. The Lee bounds range from 0.95 to 1.00 standard deviations for the Performance Task SimSE outcome, 0.62 to 0.85 for the Classroom SimSE outcome, 0.18 to 0.31 for the Classroom QCI outcome, and 0.05 to 0.26 for the Classroom MQI outcome. Because the Lee bound intervals are relatively tight and align with the main results, differential attrition does not appear to pose a substantial threat to the validity of the design.

Interpreting the Magnitude of Effects

To aid with interpretation of effect sizes, we benchmark our results against prior studies of professional development for pre-service and in-service teachers. Our prior work across five systematic replication studies, demonstrated that coaching significantly improves PSTs' instructional performance during simulated classroom

sessions, compared to PSTs who self-reflected between simulation sessions, with a meta-analytic effect of 1.34 SD (Cohen et al., 2023). In those studies, participants received a five-minute coaching session immediately following an initial simulation, and outcomes were measured in a follow-up simulation conducted directly after participants engaged in either coaching or self-reflection, using proximal measures that were tightly aligned to focal teaching skill. In the current study, the most comparable estimate is the effect of the practice-based module on PSTs' SimSE rubric scores during performance tasks (0.96 SD). The smaller effect size here may be explained by several differences between the studies, beyond the use of different samples. For example, there was a longer interval between exposure to the practice-based module and outcome measurement in the current study, whereas our earlier work measured outcomes immediately after coaching. In addition, outcomes in the current study were assessed using performance tasks rather than simulations, which may also influence the magnitude of effect estimates. Despite these differences, the current study still finds large effects comparable to those reported in our prior experimental studies of the benefits of simulations and coaching.

Prior work also shows that coaching significantly improves in-service teachers' instructional skills, with an average meta-analytic effect of 0.49 SD (Kraft et al., 2018). In Kraft et al. (2018), skills were measured using a range classroom observation rubrics such as the Classroom Assessment Scoring System (CLASS), the Early Language and Literacy Classroom Observation (ELLCO). The most comparable effects from our study the classroom observation scores on the SimSE (0.77 SD), QCI (0.36 SD), and MQI (0.26 SD) rubrics. The effect on PSTs' SimSE scores is larger than the Kraft et al. benchmark, perhaps because the practice-based module included more than just coaching supports and there was tight alignment between module content and the outcome measure. However, the effects of the practice-based module on PSTs' QCI and MQI scores are slightly smaller than the Kraft et al. benchmark, which may be because the skills and practices captured by these rubrics are not directly targeted by the intervention. The average effect size across the three classroom observation outcomes in our study is 0.46 SD. Despite clear differences in teacher populations (pre-service versus in-service), the findings presented here are comparable to the effects of coaching on instructional skills reported by Kraft and colleagues (2018).

Discussion and Implications

Given the limited duration of teacher preparation and the vast number of new teachers entering the workforce each year, it is crucial to develop portable and scalable methods for skill development, alongside a robust empirical base about the utility of such approaches. This study provides an important proof of concept for modularizing and integrating practice-based content into existing methods coursework. The intervention materials provided PSTs opportunities to engage with video examples, or representations, of the focal teaching practice of metacognitive modeling. PSTs also had opportunities to actively engage with detailed analyses, or decompositions, of those video examples. Finally, they had multiple chances to approximate metacognitive modeling in teaching simulations, while receiving in-the-moment feedback on their developing skills from coaches. The design of our practice-based intervention module reflected theory about teacher learning and the importance of "pedagogies of enactment" that allow PSTs opportunities to try out skills before they practice them with real children (Grossman et al., 2009; Lampert & Graziani, 2009). The design also followed decades of theoretical and empirical work on the importance of "deliberate practice" with feedback and chances for "do-overs" when developing new skills (Ericsson & Poole, 2016).

Despite the proliferation of theory and small-scale, qualitative studies in this space, ours is the first study to our knowledge that demonstrates the causal effect of a fully modularized, online approach to practice-based teacher education. We can make strong claims about the utility of this approach to teacher skill development due to our rigorously designed randomized control trial, coupled with a rich and varied set of outcomes collected through multiple modalities and contexts. Our data also make clear that the more typical approach to online, modularized content—reading articles and responding to prompts—will not expedite skill development to the same degree as having the chance to practice *doing* the focal skills, getting feedback from an expert coach, and then having the chance to try doing the skill again. Though the business-as-usual approach may have appeal to policymakers given low development and implementation costs, as anyone who has gone through a required e-learning module (e.g., IRB training) can attest, such experiences rarely lead to the kind of deep learning teachers need to support their work in classrooms.

At the same time, we acknowledge the practice-based module is resource intensive. It requires mixed-reality simulation software, a trained actor to voice the student avatars, and skilled coaches. These factors all increase implementation costs and may limit immediate adoption in many teacher preparation programs. We therefore view this study as evidence of the efficacy of a practice-based pedagogical approach, rather than evidence that the current delivery model is broadly affordable or readily scalable. Future research should examine whether the core features of this approach can be delivered through lower-cost formats without substantially diminishing its effects.

It is also worth noting that along with our technology partner, Mursion, we have recently developed AI-driven versions of teaching simulations in elementary mathematics. We have also been able to develop AI-driven simulation coaching experiences that mirror the model of human coaching described here. These innovations substantially drive down costs and enhance the likelihood of scaling the approach detailed here. Our research team has several studies underway to study the effects of these AI-driven simulation and coaching sessions on teacher skill development. Fields like medicine and aviation are increasingly relying on AI-driven simulations to promote skill development, but there has been little parallel innovation in education, where the needs are as great and the stakes are as high. We are excited to see whether and how teacher education can leverage these new technologies to bring the promise of practice-based learning experiences to scale.

Teacher preparation programs— and the policymakers who determine licensure and program approval requirements— desperately need evidence about flexible methods for helping support skill development within an increasingly truncated period. In the current landscape of teacher preparation, where many enter classrooms without formal student teaching, it is even more urgent to develop scalable chances to try out new skills. Ideally, less authentic approximations like simulations would serve as a complement or bridge to the more extended and realistic approximation of clinical practice (i.e. student teaching) in the ways they did for the PSTs in our study (Guha et al., 2017; Lampert et al., 2013). However, many of the online-only teacher education programs that have proliferated in recent years do not include *any* clinical practice (Bland et al., 2023; King & Yin, 2022; Kirksey & Gottlieb, 2025). A benefit of practice-based modularized content is that it can be

integrated into existing courses in university-based preparation programs, as it was in this study. But it can also serve as a standalone learning experience in other learning-to-teach contexts, including "alternative" routes, online-only programs, residencies and the like. In such program models, practice-based online modules might play an even more central role in helping PSTs learn how to enact classroom skills. In prior work, we have demonstrated the large causal effects of simulations and coaching in alternative route preparatory contexts (Cohen et al., 2025c, Wilson & Yonas, 2024). We are replicating the current study in another alternative route program to understand if the treatment effects demonstrated here are consistent in other preparatory models. It is vital that the field of teacher education research establish a strong causal evidence base about new approaches to teacher skill development across the wide range of contexts in which teachers are now prepared, and this is one step in that direction.

The rapidly shifting contexts in which teachers are prepared also suggests that existing evidence about teacher education content and pedagogy may be less applicable in the current landscape. Recent efforts to improve the accessibility of teacher education have incentivized abbreviated program durations to decrease the financial burden on PSTs (Kraft & Lyon, 2024; Tooley, 2023). As a result, time is a shrinking commodity, while the growing heterogeneity of the U.S. student population demands PSTs develop more robust skill sets for meeting the needs of diverse learners (Wood et al., 2025). Teacher educators face an increasingly complex course design task with limited resources to support their work (Hadar & Brody, 2017). This context requires creative solutions to providing high-quality, evidence-based instruction in teacher education. Our data suggest that modularizing content with practice-based supports for PSTs could be one such solution. We were able to develop and deliver a modularized learning experience that helped PSTs acquire skills known to support SWDs, who are often not the focus of instructional methods covered in general education teacher preparation. This is especially the case in elementary mathematics, where general educators are often unfamiliar with evidence-based practices for SWDs (Cohen et al., 2025b) and rarely include them in the methods courses they teach (Jones et al., 2026). These data suggest that practice-based modules developed by those with expertise in

SWDs—as well as in how PSTs learn— can provide scalable and portable avenues for expanding the repertoire of skills covered in teacher preparation.

There are some limitations to our study. All PSTs engaged with content about the focal skill of metacognitive modeling through online modules designed to take a similar amount of instructional time. We cannot make claims about the skill development of PSTs who did not engage with this content in any form. That said, we purposely created a control condition to reflect the "business-as-usual" practice in teacher education, particularly in terms of modularized content outside of standard course curricula. We often expect PSTs to learn about new skills from reading about them and reflecting on how they could be used in classrooms with students. This study makes clear that such approaches do not develop PST skills in the same way as more practice-based approaches, at least in the duration of the study. In ongoing work, we are analyzing the longer-term trajectories of skill development after engaging with a practice-based module.

An additional limitation is that these analyses were based on a sample lesson where the PSTs were prompted to demonstrate their use of metacognitive modeling in elementary mathematics, the specific instructional skill that was the focus of both the intervention and business-as-usual modules. We cannot make claims about the degree to which the observed teaching skills would be demonstrated in other, randomly selected lessons. Future research should explore whether PSTs utilize the skills featured in practice-based modules across their student teaching placements more broadly.

We also could not assess whether treatment effects varied with the time elapsed between intervention completion and outcome measurement, which limits our ability to speak to the persistence or fade-out of effects over time. Finally, as in most RCT evaluations, we cannot isolate particularly impactful components of the intervention module. The practice-based module is an intervention package, a set of interrelated learning experiences. Our data do not allow us to determine if, for example, the simulations and coaching were the aspects of the intervention that were most critical for enhancing skill development. Along the same lines, we cannot determine whether the simulations and coaching would have impacted PST skills absent the preceding work in the online learning platform. Though the "representations and decompositions" portion of the module

only took approximately three hours of time, we know that time is a very scarce commodity in teacher preparation. A multi-arm factorial design would allow us to test the impact of different combinations of "pieces" of the intervention. This would be helpful in determining the "active ingredients" that maximize the likelihood of skill development to design even more efficient interventions for teachers (Abry et al., 2011). We also have such work currently underway.

Despite these limitations, this study provides a proof of concept not only for expediting skill development, but also for new methods for evaluating the effects of teacher education interventions. Performance tasks could be a standardized and inexpensive platform for the provision of insights into teacher skill development that could provide helpful early signals of whether interventions are working in the ways we intend them (Suhr et al., In press). Importantly, however, we show that our intervention did not just lead to skill development in the short term on highly aligned outcomes. Here, we see that the practice-based module not only developed PST skills with the focal practice of metacognitive modeling, but also improved a range of moderately aligned, special education focused skills captured by the QCI, as well as the broader set of math teaching skills captured by the MQI. Teacher educators who have scarce time with PSTs want to know if new approaches or content lead to development in the longer term, in real classrooms with real children. Given the limited time in the pre-service period, they rightly also want to know if such approaches promote growth across the wide array of skills beginning teachers need. These data do not allow us to understand the mechanisms underlying the transfer of skills targeted in the intervention to those assessed by these less aligned measures, though we have additional qualitative analyses underway to help us better discern why we might have seen these results. Being able to analyze the effects of interventions across a range of outcomes in a range of contexts is essential as policymakers try to understand what experiences should be required in the limited time teacher educators have to prepare PSTs.

References

Abry, T., Hulleman, C. S., & Rimm-Kaufman, S. E. (2015). Using indices of fidelity to intervention core components to identify program active ingredients. *American Journal of Evaluation*, *36*(3), 320-338. https://doi.org/10.1177/1098214014557009

Aguilar, J. J., & Flores, Y. (2022). Analyzing the effectiveness of using mixed-reality simulations to develop elementary pre-service teacher's high-leverage practices in a mathematics methods course. *Eurasia Journal of Mathematics, Science and Technology Education*, *18*(5), 1-12. https://doi.org/10.29333/ejmste/12006

Alves, K. D., Kennedy, M. J., Kellems, R. O., Wexler, J., Rodgers, W. J., Romig, J. E., & Peeples, K. N. (2018). Improving preservice teacher vocabulary instruction: A randomized controlled trial. *Teacher Education and Special Education*, *41*(4), 340-356. https://doi.org/10.1177/0888406417727044

Ball, D., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education*, *60*(5), 497–511. https://doi.org/10.1177/0022487109348479

Banks, B., Sims, S., Curran, J., Meliss, S., Chowdhury, N., Altunbas, H.G., Alexandri, N., MacTavish, L., & Instone, I. (2024). *Decomposition and recomposition in teacher education* (No. 24-08). UCL Centre for Education Policy and Equalising Opportunities.

Bartanen, B., & Kwok, A. (2021). Examining clinical teaching observation scores as a measure of preservice teacher quality. *American Educational Research Journal, 58*(5), 887-920. https://doi.org/10.3102/0002831221990359

Bland, J. A., Wojcikiewicz, S. K., Darling-Hammond, L., & Wei, W. (2023). Strengthening pathways into the teaching profession in Texas: Challenges and opportunities. *Learning Policy Institute.* https://files.eric.ed.gov/fulltext/ED630218.pdf

Boguslav, A., & Cohen, J. (2024). Different Methods for Assessing Preservice Teachers' Instruction: Why Measures Matter. *Journal of Teacher Education*, *75*(2), 168-185.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brownell, M. T., Jones, N. D., Sohn, H., & Stark, K. (2020). Improving teaching quality for students with disabilities: Establishing a warrant for teacher education practice. *Teacher Education and Special Education, 43*(1), 28-44. https://doi.org/10.1177/0888406419880351

Brownell, M. T., Sindelar, P. T., Kiely, M. T., & Danielson, L. C. (2010). Special education teacher quality and preparation: Exposing foundations, constructing a new model. *Exceptional Children*, *76*(3), 357-377.

Carpenter, T. P., Hiebert, J., Fennema, E., Fuson, K., Wearne, D., & Murray, H. (1997). *Making sense: Teaching and learning mathematics with understanding*. Heinemann. https://www.heinemann.com/products/ebk03022.aspx

Carver-Thomas, D., Burns, D., Leung, M., & Ondrasek, N. (2022). Teacher shortages during the pandemic: How California districts are responding. *Learning Policy Institute*.

Clark, L. M., DePiper, J. N., Frank, T. J., Nishio, M., Campbell, P. F., Smith, T. M., Griffin, M. J., Rust, A. H., Conant, D. L. & Choi, Y. (2014). Teacher characteristics associated with mathematics teachers' beliefs and awareness of their students' mathematical dispositions. *Journal for Research in Mathematics Education, 45*(2), 246-284.

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). *Teacher coaching in a simulated environment*. *Educational Evaluation and Policy Analysis, 42*(2), 208–231. https://doi.org/10.3102/0162373720906217

Cohen, J., Wong, V. C., Krishnamachari, A., & Erickson, S. (2023). *Experimental evidence on the robustness of coaching supports in teacher education*. *Educational Researcher, 53*(1), 19–35. https://doi.org/10.3102/0013189X231198827

Cohen, J., Yonas, A., & Wilson, K. (2025a). *Approximating teaching: A systematic review of the literature*. *Review of Educational Research*. https://doi.org/10.3102/00346543251368383

Cohen, J., Jones, N., & Gibbons, L. (2025b). The Missing Middle? General and Special Educators' Views of Effective Mathematics Instruction. *AERA Open*, *11*.

Cohen, J., Wong, V., Liu, Q., Wilson, K., Yonas, A., & Parten, H. (2025c). *Practice does not make perfect: Experimental evidence on the effectiveness of coaching beginning teachers*. *Journal of Educational Psychology, 117*(7), 1137–1177. https://doi.org/10.1037/edu0000960

Davis, E. A., Kloser, M., Wells, A., Windschitl, M., Carlson, J., & Marino, J. C. (2017). Teaching the practice of leading sense-making discussions in science: Science teacher educators using rehearsals. *Journal of Science Teacher Education*, *28*(3), 275-293. https://doi.org/10.1080/1046560X.2017.1302729

Dawson, H., & Scott, L. (2013). Teaching students with disabilities efficacy scale: Development and validation. *Inclusion, 1*(3), 181-196. https://doi.org/10.1352/2326-6988-1.3.181

Dieker, L. A., Rodriguez, J. A., Lignugaris, Kraft, B., Hynes, M. C., & Hughes, C. E. (2014). The potential of simulated environments in teacher education: Current and future possibilities. *Teacher Education and Special Education*, *37*(1), 21–33.

Doabler, C. T., Nelson, N. J., Kosty, D. B., Fien, H., Baker, S. K., Smolkowski, K., & Clarke, B. (2014). Examining teachers' use of evidence-based practices during core mathematics instruction. *Assessment for Effective Intervention*, *39*(2), 99-111.

Dutro, E., & Cartun, A. (2016). Cut to the core practices: Toward visceral disruptions of binaries in PRACTICE-based teacher education. *Teaching and Teacher Education*, *58*, 119-128. https://doi.org/10.1016/j.tate.2016.05.001

Driver, M. K., Zimmer, K. E., Khan, O., Sadler, J. V., & Draper, E. (2024). Preparing general education teachers for inclusive Settings: Integrating high-leverage practices and mixed-reality simulation in pre-service coursework. *Education Sciences*, *14*(4), 428. https://doi.org/10.3390/educsci14040428

Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the new science of expertise.* Houghton Mifflin Harcourt.

García, A. I., Jiménez, J. E., & Hess, S. (2006). Solving arithmetic word problems: An analysis of classification as a function of difficulty in children with and without arithmetic LD. *Journal of Learning Disabilities*, *39*(3), 270–281. https://doi.org/10.1177/00222194060390030601

Goering S. (2015). Rethinking disability: the social model of disability and chronic disease. *Current Reviews in Musculoskeletal Medicine*, *8*(2), 134–138. https://doi.org/10.1007/s12178-015-9273-z

Goldhaber, D., Krieg, J., Naito, N., & Theobald, R. (2020). Making the most of student teaching: The importance of mentors and scope for change. *Education Finance and Policy*, *15*(3), 581–591. https://doi.org/10.1162/edfp_a_00305

Grossman, P. (2008). Responding to our critics: From crisis to opportunity in research on teacher education. *Journal of Teacher Education, 59*(1), 10-23. https://doi.org/10.1177/0022487107310748

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), 2055-2100.

Grossman, P., Kaul, M., & Kavanagh, S. S. (2026). Landscape analysis of the teaching profession. EdWorkingPaper: 26-1399. *Annenberg Institute at Brown University*. https://doi.org/10.26300/d4g1-hy54

Guha, R., Hyler, M. E., & Darling-Hammond, L. (2017). The teacher residency: A practical path to recruitment and retention. *American Educator, 41*(1), 31-34. https://files.eric.ed.gov/fulltext/EJ1137804.pdf

Hadar, L., & Brody, D. (2017). Professional learning and development of teacher educators. In D. J. Clandinin & J. Husu (Eds.), *The SAGE Handbook of Research on Teacher Education Vol. 2* (pp. 1049-1064). SAGE Publications. https://doi.org/10.4135/9781526402042.n60

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831. https://doi.org/10.3102/0002831210387916

Hill, H. C., Mancenido, Z., & Loeb, S. (2024). Effectiveness research for teacher education. *Educational Researcher*, 53(6), 370-377. https://doi.org/10.3102/0013189X241260393

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal, 105(*1), 11-30.

James, M. C., & Scharmann, L. C. (2007). Using analogies to improve the teaching performance of preservice teachers. *Journal of Research in Science Teaching*, *44*(4), 565–585.

Jay, J. K., & Johnson, K. L. (2002). Capturing complexity: A typology of reflective practice for teacher education. *Teaching and teacher education*, *18*(1), 73-85.

Jones, N. D., Cohen, J., Shaheen, T., Ristroph, I., Sebastian, R., & Aigotti, S. (2026). Teacher Candidate Preparation for Inclusion: Coverage of Students with Disabilities in Mathematics Methods Syllabi. *Journal of Teacher Education*, *77*(3), 197-211.

Judge, S., & Watson, S. M. (2011). Longitudinal outcomes for mathematics achievement for students with learning disabilities. *The Journal of Educational Research*, *104*(3), 147-157.

Kavanagh, S. S., & Danielson, K. A. (2020). Practicing justice, justifying practice: Toward critical practice teacher education. *American Educational Research Journal*, *57*(1), 69-105. https://doi.org/10.3102/0002831219848691

Kennedy, M. M. (1999). The role of preservice teacher education. In L. Darling Hammond & G. Sykes (Eds.), *Teaching as the Learning Profession: Handbook of Policy and Practice* (pp. 54–85). Jossey Bass.

Kennedy, M. M. (2016). Parsing the practice of teaching. *Journal of Teacher Education*, *67*(1), 6–17. https://doi.org/10.1177/0022487115614617

Kennedy, M. J., & Romig, J. E. (2021). Cognitive Load Theory: An applied reintroduction for special and general educators. *Teaching Exceptional Children, 56(6), 440-451.* https://doi.org/10.1177/00400599211048214

Kennedy, M. J., Hart, J. E., & Kellems, R. O. (2011). Using enhanced podcasts to augment limited instructional time in teacher preparation. *Teacher Education and Special Education, 34*(2), 87-105. https://doi.org/10.1177/0888406410376203

Kennedy, M. J., Hirsch, S. E., Dillon, S. E., Rabideaux, L., Alves, K. D., & Driver, M. K. (2016). Using content acquisition podcasts to increase student knowledge and to reduce perceived cognitive load. *Teaching of Psychology, 43*(2), 153-158. https://doi.org/10.1177/0098628316636295

King, J. E., & Yin, J. (2022). The alternative teacher certification sector outside higher education: 2022 Update. *Center for American Progress.* https://files.eric.ed.gov/fulltext/ED626078.pdf

Kirksey, J. J., & Gottlieb, J. J. (2025). Virtual pathways: Linking online alternative certification programs to student achievement and teacher attrition in Texas. *Educational Researcher, 55*(3). https://doi.org/10.3102/0013189X251386406

Kraft, M.A., Blazar, D., Hogan, D. (2018). The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research,* 88(4), 547-588.

Kraft, M. A., & Lyon, M. A. (2024). The rise and fall of the teaching profession: Prestige, interest, preparation, and satisfaction over the last half century. *American Educational Research Journal, 61*(6), 1192-1236. https://doi.org/10.3102/00028312241276856

Lampert, M., Franke, M. L., Kazemi, E., Ghousseini, H., Turrou, A. C., Beasley, H., Cunard, A. & Crowe, K. (2013). Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education, 64*(3), 226-243. https://doi.org/10.1177/0022487112473837

Lampert, M., & Graziani, F. (2009). Instructional Activities as a Tool for Teachers' and Teacher Educators' Learning. *The Elementary School Journal*, *109*(5), 491–509. https://doi.org/10.1086/596998

Lautenbach, F., & Heyder, A. (2019). Changing attitudes to inclusion in preservice teacher education: a systematic review. *Educational Research*, *61*(2), 231–253. https://doi.org/10.1080/00131881.2019.1596035

Leko, M. M., Brownell, M. T., O'Brien, K. & Nagro, S. A. (2023). Capitalizing on the potential of practice based preparation in special education teacher preparation. In E.D. McCray, E. Bettini, M.T. Brownell, J. McLeskey & P.T. Sindelar (Eds.), *Handbook of Research on Special Education Teacher Preparation (pp. 106 – 125).* Routledge. https://doi.org/10.4324/9781003297093-8

Lewis, K. E., & Lynn, D. M. (2018). Against the odds: Insights from a statistician with dyscalculia. *Education Sciences*, *8*(2), 63. https://doi.org/10.3390/educsci8020063

Lein, A. E., Jitendra, A. K., & Harwell, M. R. (2020). Effectiveness of mathematical word problem solving interventions for students with learning disabilities and/or mathematics difficulties: A meta-analysis. *Journal of Educational Psychology, 112(*7), 1388-1408. https://doi.org/10.1037/edu0000453

Mancenido, Z. (2024). Impact evaluations of teacher preparation practices: Challenges and opportunities for more rigorous research. *Review of Educational Research*, *94*(2), 268-307. https://doi.org/10.3102/00346543231174413

Mancenido, Z., Hill, H. C., Garcia Coppersmith, J., Carter, H., Pollard, C., & Monschauer, C. (2026). Practice-Based teacher education pedagogies improve responsiveness: Evidence from a lab experiment. *Journal of Research on Educational Effectiveness*, *19*(1), 201–225. https://doi.org/10.1080/19345747.2025.2456716

Mayer R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press.

McLeskey, J., Barringer, M-D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M., Lewis, T., Maheady, L., Rodriguez, J., Scheeler, M. C., Winn, J., & Ziegler, D. (2017). *High leverage practices in special education*. Council for Exceptional Children & Collaboration for Effective Educator Development, Accountability and Reform (CEEDAR). https://ceedar.education.ufl.edu/wp-content/uploads/2017/07/CEC-HLP-Web.pdf

Montague, M. (1992). The effects of cognitive and metacognitive strategy instruction on the mathematical problem solving of middle school students with learning disabilities. *Journal of learning disabilities*, *25*(4), 230-248.

Montague, M., & Dietz, S. (2009). Evaluating the evidence base for cognitive strategy instruction and mathematical problem solving. *Exceptional Children, 75(*3), 285–302. https://doi.org/10.1177/001440290907500302

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, *91*(2), 358. https://doi.org/10.1037/0022-0663.91.2.358

Moreno, R., & Valdez, A. (2007). Immediate and delayed effects of using a classroom case exemplar in teacher education: The role of presentation format. *Journal of Educational Psychology, 99*(1), 194–206. https://doi.org/10.1037/0022-0663.99.1.194

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, *108*(4), 455–473. https://doi.org/10.1037/edu0000079

Peng, P., Wang, C., & Namkung, J. (2018). Understanding the cognition related to mathematics difficulties: A meta-analysis on the cognitive deficit profiles and the Bottleneck Theory. *Review of Educational Research*, *88*(3), 434–476. https://doi.org/10.3102/0034654317753350

Phelps, G., Kinsey, D., Florek, T., & Jones, N. (2023). Using performance tasks to provide feedback and assess progress in teacher preparation. *ETS Research Report Series,* 2023(1), 1-24. https://doi.org/10.1002/ets2.12371

Philip, T. M., Souto-Manning, M., Anderson, L., Horn, I., Andrews, D. J. C., Stillman, J., & Varghese, M. (2019). Making justice peripheral by constructing practices as “core:” How the increasing prominence of core practices challenges teacher education. *Journal of Teacher Education*, 70(3), 251-264.

Powell, S. R., & Fuchs, L. S. (2018). Effective word-problem instruction: Using schemas to facilitate mathematical reasoning. *Teaching Exceptional Children, 51*(1), 31-42. https://doi.org/10.1177/0040059918777250

Powell, S. R., Berry, K. A., & Benz, S. A. (2020). Analyzing the word-problem performance and strategies of students experiencing mathematics difficulty. *Journal of Mathematical Behavior*, *58.* https://doi.org/10.1016/j.jmathb.2020.100759

Quigley, A., Muijs, D., & Stringer, E. (2018). Metacognition and self-regulated learning: Guidance report. *Education Endowment Foundation.*

Reid, R., Lienemann, T. O., & Hagaman, J. L. (2013). *Strategy instruction for students with learning disabilities*, Second Edition. Guilford Publications.

Sayeski, K. L., Hamilton-Jones, B., & Oh, S. (2015). The efficacy of IRIS *STAR Legacy* Modules under different instructional conditions. *Teacher Education and Special Education, 38*(4), 291–305.

Schulte, A. C., & Stevens, J. J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children*, *81*(3), 370-387.

Schutz, K. M., Grossman, P., & Shaughnessy, M. (2018). Approximations of practice in teacher education. In P. Grossman (Ed.), *Teaching Core Practices in Teacher Education* (pp. 57 - 83). Harvard Education Press

Smith, D. D., Lewis, J., Montrosse-Moorhead, B. M., & Brown, J. (2014). *Use of IRIS Centers' resources at institutions of higher education with approved special education licensure programs: 2013-2014 academic year*. Claremont, CA: IRIS Center West, Claremont Graduate University.

South Carolina Department of Education. (2025, August 18). *2016 dyslexia training as required by Joint Resolution 5024.* https://ed.sc.gov/districts-schools/special-education-services/additional-information-and-assistance/dyslexia-and-other-reading-disorders/dyslexia-final-report-pdf/

Stroupe, D., & Gotwals, A. W. (2018). "It's 1000 degrees in here when I teach": Providing preservice teachers with an extended opportunity to approximate ambitious instruction. *Journal of Teacher Education*, *69*(3), 294-306.

Suhr, M., Erickson, S., Kelcey, B., McLean, L., Cohen, J., & Jones, N. (In press). Exploring the use of performance tasks as formative assessments of pre-service teacher instructional skills within teacher preparation programs. *Educational Assessment.*

Tooley, M. (2023). *National scan of pathways to becoming a first-time teacher* (Report No. ED627948). New America. https://files.eric.ed.gov/fulltext/ED627948.pdf

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. (2018). *Improving Mathematical Problem Solving in Grades 4 Through 8* (NCEE 2012-4055). Retrieved from What Works Clearinghouse

Virginia Department of Education. (2025, August 18). *Dyslexia training.* https://www.doe.virginia.gov/teaching-learning-assessment/teaching-in-virginia/teacher-licensure/dyslexia-training

Wei, X., Lenz, K. B. & Blackorby, J. (2013). Math growth trajectories of students with disabilities: disability category, gender, racial, and socioeconomic status differences from ages 7 to 17. *Remedial and Special Education, 34*(3), pp. 154-165.

Wilson, K., & Yonas, A. M. (2024). In search of deliberate practice: Simulating teaching in three teacher education programs. *Teachers College Record, 126*(9), 47-89. https://doi.org/10.1177/01614681241298842

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2001). Teacher preparation research: Current knowledge, gaps, and recommendations. *Center for the Study of Teaching and Policy.* https://www.researchgate.net/publication/242434276_Teacher_Preparation_Research_Current_Knowledge_Gaps_and_Recommendations

Wood, W. J., Lai, I., Filosa, N. R., Imberman, S. A., Jones, N. D., & Strunk, K. O. (2025). Are effective teachers for students with disabilities effective teachers for all? *Educational Evaluation and Policy Analysis, 47*(2), 355–379. https://doi.org/10.3102/01623737231214555

Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A.,Koedinger, K.R.& Ogbuehi, P. (2012). Improving mathematical problem solving in grades 4 through 8. IES Practice Guide. NCEE 2012-4055. *What Works Clearinghouse*.

Zeichner, K. (2012). The turn once again toward practice-based teacher education. *Journal of Teacher Education*, *63*(5), 376-382. https://doi.org/10.1177/0022487112445789

Table 1. Sample Description

	Overall	Site 1 Fall	Site 1 Spring	Site 2 Fall	Site 2 Spring
% White	0.78	0.84	0.74	0.73	0.82
% Male	0.01	0.00	0.04	0.00	0.00
Age	20.82	24.74	20.74	20.10	20.14
College GPA	3.59	3.56	3.64	3.67	3.49
SimSe Pretest	1.22	1.32	1.24	1.20	1.21
Math Knowledge for Teaching	0.00	0.25	-0.37	0.30	-0.20
Self Efficacy	3.45	3.20	3.49	3.55	3.42
Beliefs about Teaching	2.85	2.87	2.66	2.95	2.83
Prior Experience	2.16	2.16	2.16	2.17	2.14
N	149	19	27	52	51

Table 2. Baseline Balance of Experimental Sample

	Control Mean	Treatment Mean	Difference
% White	0.78	0.78	-0.00
			(0.95)
% Male	-0.00	0.01	0.01
			(0.33)
Age	20.66	20.97	0.32
			(0.62)
College GPA	3.54	3.63	0.09
			(0.43)
SimSe Pretest	1.21	1.23	0.02
			(0.55)
Math Knowledge for Teaching	-0.02	0.02	0.05
			(0.77)
Self Efficacy	3.39	3.50	0.11
			(0.25)
Beliefs about Teaching	2.85	2.85	0.00
			(0.97)
Prior Experience	2.21	2.11	-0.10
			(0.20)

The differences reported in column three are estimated by regressing each covariate on an indicator for treatment status, controlling for randomization blocks. The P-values for the estimated differences are reported underneath in parenthesis. We also used a joint F-test based on a multivariate regression of all covariates on treatment status and block indicators. The test statistic is $F(9, 144) = 0.45$, with a p-value of 0.9023.

Table 3. Standardized Intent-to-Treat Estimates Across Multiple Measures

	(1) Performance Task SimSE Coef./SE	(2) Classroom SimSE Coef./SE	(3) Classroom QCI Coef./SE	(4) Classroom MQI Coef./SE
ITT	0.961^{***}	0.769^{***}	0.362^{***}	0.261^{*}
	(0.11)	(0.12)	(0.12)	(0.14)
N. Observations	860	358	484	512
N. PSTs	146	139	139	140
Rater FE	X	X	X	X
Segment FE		X	X	X
Task FE	X			
Clustered SEs	X	X	X	X

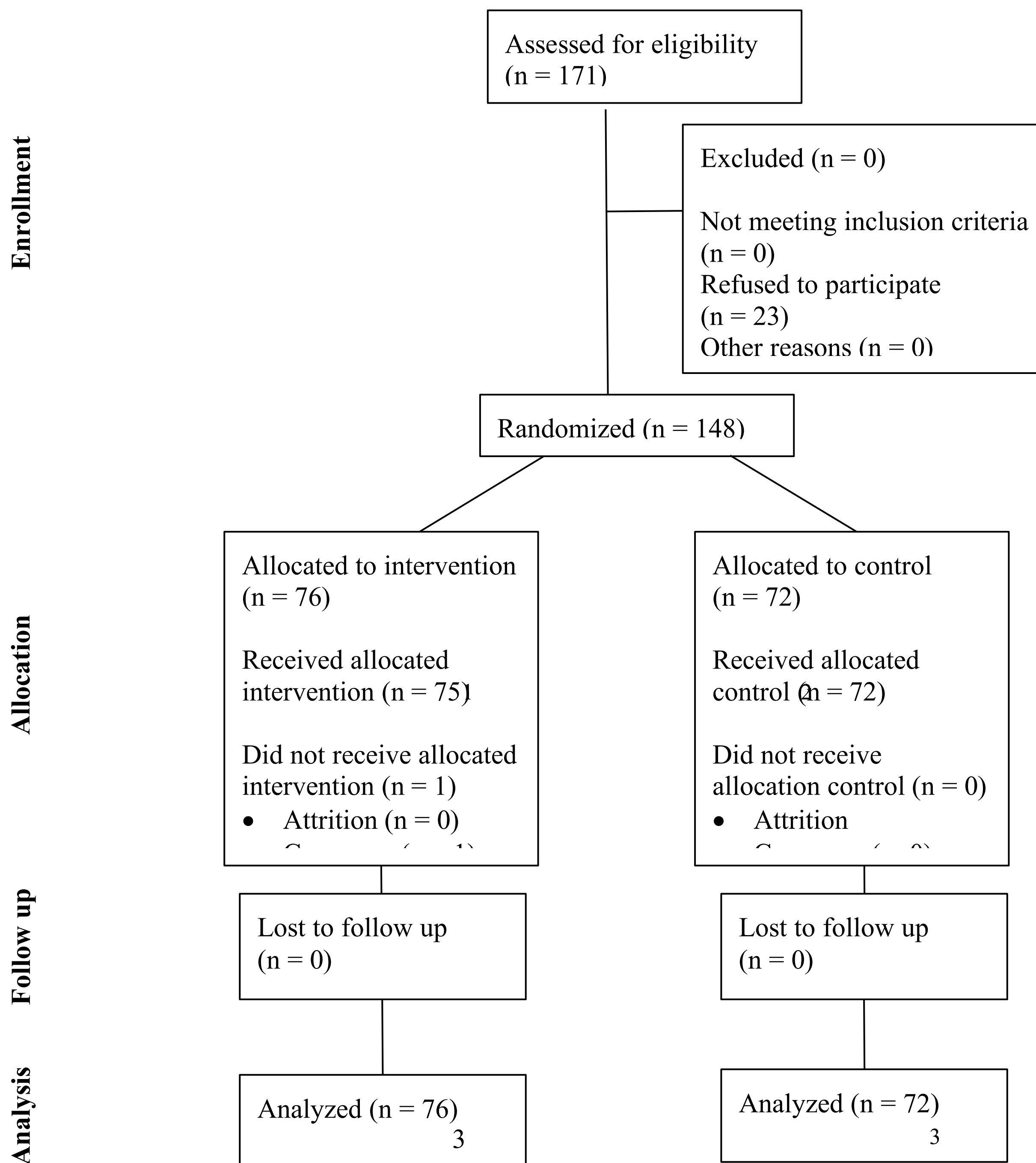
Note: Table reports standardized ITT effects of the intervention on four outcome measures of instructional quality. The coefficients are estimated using the model in equation 1. Standard errors are clustered at the PST level to account for repeated observations. Columns 1 and 2 present effects on metacognitive modeling scored using the SimSE rubric, drawn from standardized performance tasks and classroom videos, respectively. Columns 3 and 4 present effects on broader measures of instructional quality: the Quality of Classroom Instruction (QCI) and the Mathematical Quality of Instruction (MQI). SimSE and QCI rubrics are scored on a 3-point scale ranging from 1 ("low quality") to 3 ("high quality"); MQI lesson segment items are scored on a 0–3 scale and MQI whole lesson items are scored on a 1-5 scale, with higher scores indicating stronger mathematics instruction. All models include fixed effects for raters, segments (or tasks), and randomization blocks, and cluster standard errors at the PST level. $p < .10$, p < .05, $p < .01$.

Table 4. Outcome Measures Descriptive Statistics for the Control Sample

		Mean	SD	Min	Max
Performance Task	SimSE	1.31	0.22	1.00	2.03
Classroom Tasks	SimSE	1.52	0.32	1.00	2.67
	QCI	2.35	0.40	1.12	3.00
	MQI	1.87	0.27	1.13	2.51

Note: SimSE and QCI scores range from 1 (“low quality”) to 3 (“high quality”); MQI lesson segment scores range from 0 (“not present”) to 3 (“high quality”); MQI whole lesson scores range from 1 (“not at all true of this lesson”) to 5 (“very true of this lesson”), with intermediate values reflecting increasing levels of instructional quality.

Figure 1. Consort Diagram



- Adherence for intervention group: Completed asynchronous module: n = 74; Completed both simulations: n = 69
- Adherence for control group: Completed all reflection: n = 66
- Numbers differ by outcome; please see Table C1

Table C1. N Analyzed Per Arm For Each Outcome

		N	Reason for Missingness
SimeSE Performance Task	Business-as-Usual	71	• No video submission (n=1)
	Treatment	75	• No video submission (n=1)
SimeSE Classroom	Business-as-Usual	66	• No video submission (n=2) • Video unable to be scored (n=4)
	Treatment	73	• No video submission (n=1) • Video unable to be scored (n=2)
MQI Classroom	Business-as-Usual	66	• No video submission (n=2) • Video unable to be scored (n=4)
	Treatment	74	• No video submission (n=1) • Video unable to be scored (n=1)
QCI Classroom	Business-as-Usual	66	• No video submission (n=2) • Video unable to be scored (n=4)
	Treatment	73	• No video submission (n=1) • Video unable to be scored (n=1)

Figure 2. Examples of Interactive Features

What Does Metacognitive Modeling Sound Like?

Click each of the boxes below for examples of metacognitive modeling while unpacking word problems:

Examples of Self-Questioning

Note: These statements are general, and can be applied to various problems. This supports students using self-questioning to make sense of other problems. The answers will be specific.

"What is this problem about?" ... "This problem is about... I know this because...."

"What are some ways we have solved problems like this before? This can help us think about some approaches that might work."

"What might this answer sound like?" ... "I think this answer might sound like '*Blank number of people are ...*'"

The teacher illustrates both an important question learners can ask themselves when first attempting to make sense of a problem and important details they pulled from the word

The teacher illustrates that when deciding on a problem solving approach, learners might need to think about several possible approaches and choose the one that they think

The teacher illustrates that a learner might not know an answer yet, but thinking about how the answer could sound can help clarify the context of the problem. This can help them determine a

Example or Non-Example?

Based on what you've learned in this module so far, click and drag to sort the statements below into examples and non-examples.

"The first thing I do is always circle the numbers in the problem."

"I'm not sure what to do, so I'll re-read to see what the problem is asking."

Example

Non-example

DEBRIEF THE SORT

Figure 3. A PST Enacts a Metacognitive Model

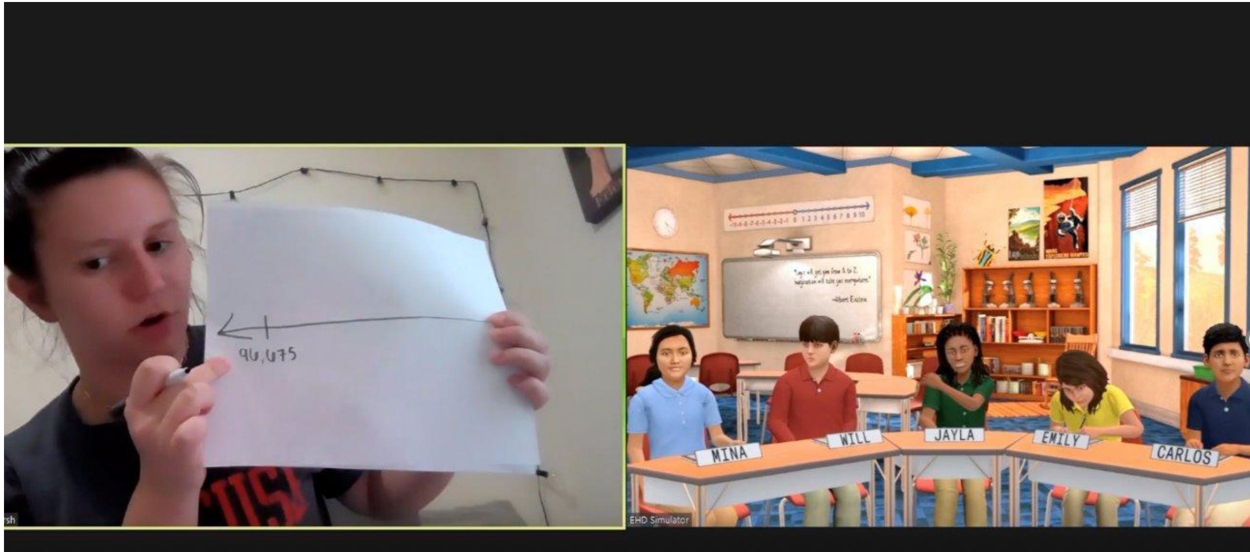


Figure 4. Module Design: Practice-Based versus Reading and Reflection

Figure 5. Overview of Study Data Collection Procedures

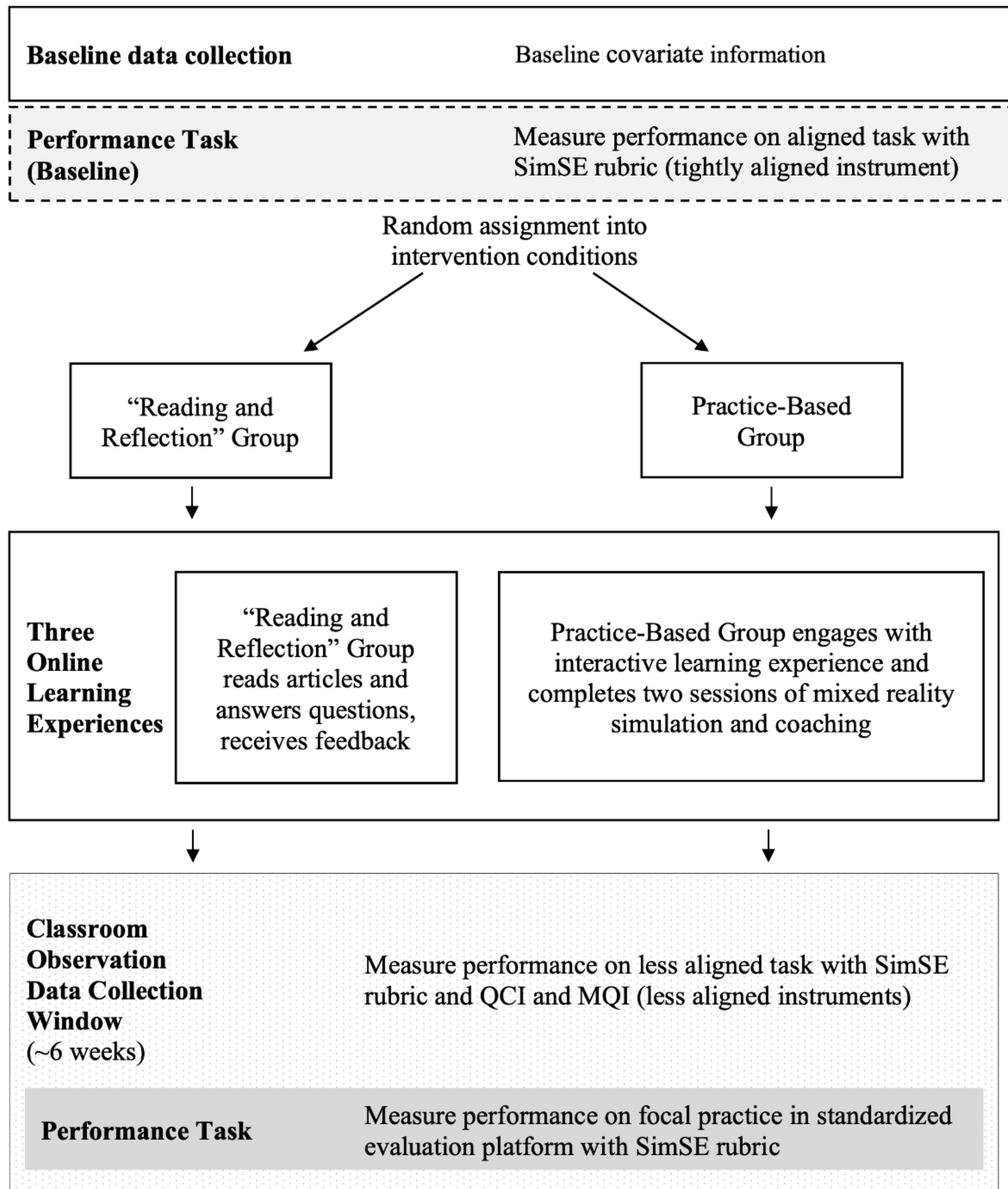


Figure 6. Performance Task Data Collection Platform

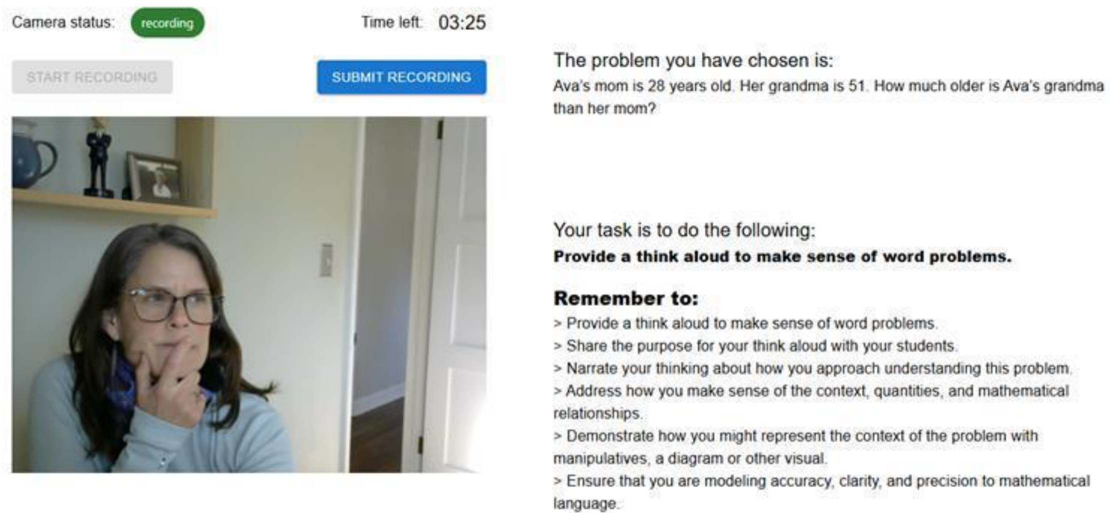
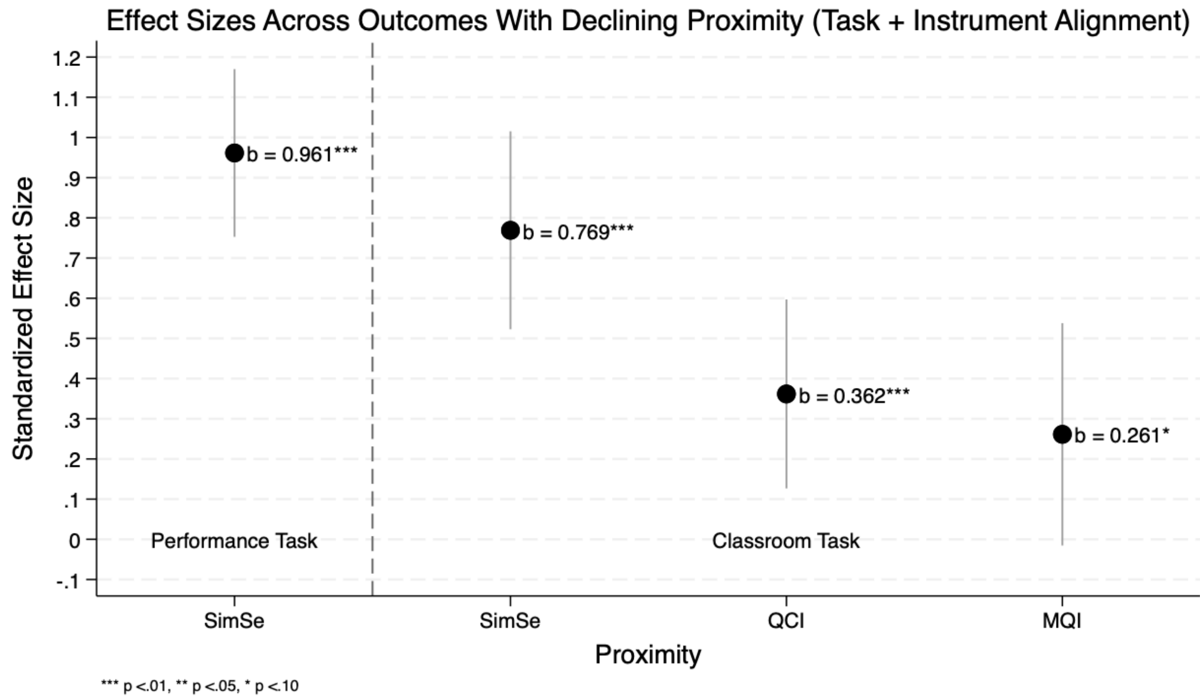


Figure 7. Standardized Effects of Practice-Based Module as Compared to Reading and Reflection Module Across Outcome Measures



Appendix Tables

Table A1. *Standardized TOT Effects Across Multiple Measures*

	(1) Performance Task SimSE Coef./SE	(2) Classroom SimSE Coef./SE	(3) Classroom QCI Coef./SE	(4) Classroom MQI Coef./SE
TOT	0.977***	0.798***	0.366***	0.264*
	(0.10)	(0.12)	(0.12)	(0.14)
N. Observations	860	358	484	512
N. PSTs	146	139	139	140
Rater FE	X	X	X	X
Segment FE		X	X	X
Task FE	X			
Clustered SEs	X	X	X	X

Note: Table reports standardized TOT effects of the intervention on four outcome measures of instructional quality. The coefficients are estimated using the model in equation 1. Standard errors are clustered at the PST level to account for repeated observations. Columns 1 and 2 present effects on metacognitive modeling scored using the SimSE rubric, drawn from standardized performance tasks and classroom videos, respectively. Columns 3 and 4 present effects on broader measures of instructional quality: the Quality of Classroom Instruction (QCI) and the Mathematical Quality of Instruction (MQI). SimSE and QCI rubrics are scored on a 3-point scale ranging from 1 ("low quality") to 3 ("high quality"); MQI lesson segment scores range from 0 ("not present") to 3 ("high quality"); MQI whole lesson scores range from 1 ("not at all true of this lesson") to 5 ("very true of this lesson"). All models include fixed effects for raters, segments (or tasks), and randomization blocks, and cluster standard errors at the PST level. $p < .10$, p < .05, $p < .01$.

Table A2. Balance Tables for Baseline and Analytic Samples

	1 Baseline (n = 149)		2 Performance Assessment (n = 146)		3 Classroom Observation (n = 139)				
	Control	Treatment	Control	Treatment	Control	Treatment	Difference 1	Difference 2	Difference 3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
% White	0.78	0.78	0.79	0.77	0.77	0.77	-0.00	-0.01	-0.01
							(0.95)	(0.66)	(0.94)
% Male	-0.00	0.01	0.00	0.01	-0.00	0.01	0.01	0.01	0.01
							(0.33)	(0.02)	(0.34)
Age	20.66	20.97	20.64	20.94	20.32	21.01	0.32	0.30	0.70
							(0.62)	(0.26)	(0.26)
College GPA	3.54	3.63	3.56	3.65	3.54	3.62	0.09	0.09	0.07
							(0.43)	(0.05)	(0.52)
SimSe Pretest	1.21	1.23	1.23	1.24	1.23	1.23	0.02	0.01	0.00
							(0.55)	(0.48)	(0.90)
Math Knowledge for Teaching	-0.02	0.02	-0.03	0.03	0.06	0.15	0.05	0.06	0.09
							(0.77)	(0.35)	(0.60)
Self Efficacy	3.39	3.50	3.39	3.50	3.37	3.51	0.11	0.11	0.13
							(0.25)	(0.01)	(0.20)
Beliefs about Teaching	2.85	2.85	2.85	2.87	2.86	2.84	0.00	0.02	-0.02
							(0.97)	(0.55)	(0.82)
Prior Experience	2.21	2.11	2.20	2.10	2.20	2.12	-0.10	-0.11	-0.08
							(0.20)	(0.00)	(0.37)

The table reports descriptive statistics for the baseline sample and two analytic samples: those with performance assessment data and those with classroom observation data. Columns 1–6 show means for each covariate by treatment status. Columns 7–9 report the difference in means between treatment and control groups for each sample. P-values of the differences, shown in parentheses, are from regressions of each covariate on treatment status with randomization block fixed effects.

Table A3. Survey Measure Reliability

Survey Measure	Number of items	Cronbach's Alpha
Math Knowledge for Teaching	37	0.7559
Self-Efficacy	19	0.9326
Beliefs about Teaching	21	0.7032
Prior Experience	31	0.6815
NEO Personality	60	0.8642

The table reports Cronbach’s alpha as a measure of reliability for each survey measure used as a covariate in the regression models used to estimate treatment effects

Table A4. Outcome Measure Reliability

Outcome	Error Variance	True Score Variance	Reliability
Performance Task Meta Model	0.02	0.08	0.80
Classroom Meta Model	0.05	0.09	0.66
Classroom Task QCI	0.11	0.04	0.27
Classroom Task MQI	0.05	0.02	0.33

Reliability is measured using Generalizability Coefficients. Generalizability Coefficients were calculated using a rater-nested-within-participant design. We treated raters as nested because each participant was assigned a different combination of two raters (except for the MQI outcome, where the same two raters double-scored all participants). The Generalizability Coefficient was computed as the proportion of between-participant (true score) variance relative to total variance (true score variance plus rater error variance). The reported Generalizability Coefficients therefore describe the reliability of a single score assigned by one rater to a participant. The outcome scores, as they are used in the regression models, are slightly more reliable because we have one score from two distinct raters, which decreases the error variance by $\frac{Error\ variance}{2}$

Table A5. Standardized ITT Effects Across Multiple Measures Without Covariates

	(1) Performance Task SimSE Coef./SE	(2) Classroom SimSE Coef./SE	(3) Classroom QCI Coef./SE	(4) Classroom MQI Coef./SE
ITT	0.993^{***}	0.666^{***}	0.273^{**}	0.151
	(0.13)	(0.13)	(0.14)	(0.14)
N. Observations	860	358	484	512
N. Candidates	146	139	139	140
Rater FE	X	X	X	X
Segment FE		X	X	X
Task FE	X			
Clustered SEs	X	X	X	X

Note: Table reports standardized TOT effects of the intervention on four outcome measures of instructional quality. The coefficients are estimated using the model in equation 1 with the vector of baseline covariates removed from the model. Standard errors are clustered at the PST level to account for repeated observations. Columns 1 and 2 present effects on metacognitive modeling scored using the SimSE rubric, drawn from standardized performance tasks and classroom videos, respectively. Columns 3 and 4 present effects on broader measures of instructional quality: the Quality of Classroom Instruction (QCI) and the Mathematical Quality of Instruction (MQI). SimSE and QCI rubrics are scored on a 3-point scale ranging from 1 (“low quality”) to 3 (“high quality”); MQI lesson segment scores range from 0 (“not present”) to 3 (“high quality”); MQI whole lesson scores range from 1 (“not at all true of this lesson”) to 5 (“very true of this lesson”). All models include fixed effects for raters, segments (or tasks), and randomization blocks, and cluster standard errors at the PST level. $p < .10$, p $< .05$, $p < .01$.

Table A6. Standardized ITT Effects Across Multiple Measures Without Covariates or Rater and Segment Fixed Effects

	(1) Performance SimSE Coef./SE	(2) Classroom SimSE Coef./SE	(3) Classroom QCI Coef./SE	(4) Classroom MQI Coef./SE
ITT	0.992***	0.686***	0.198	0.164
	(0.13)	(0.15)	(0.14)	(0.14)
N. Observations	861	358	484	512
N. Candidates	146	139	139	140
Rater FE	X	X	X	X
Segment FE		X	X	X
Task FE	X			
Clustered SEs	X	X	X	X

Note: Table reports standardized TOT effects of the intervention on four outcome measures of instructional quality. The coefficients are estimated using the model in equation 1 with the vector of baseline covariates and the rater and segment (or task) fixed effects removed from the model. Standard errors are clustered at the PST level to account for repeated observations. Columns 1 and 2 present effects on metacognitive modeling scored using the SimSE rubric, drawn from standardized performance tasks and classroom videos, respectively. Columns 3 and 4 present effects on broader measures of instructional quality: the Quality of Classroom Instruction (QCI) and the Mathematical Quality of Instruction (MQI). SimSE and QCI rubrics are scored on a 3-point scale ranging from 1 ("low quality") to 3 ("high quality"); MQI lesson segment scores range from 0 ("not present") to 3 ("high quality"); MQI whole lesson scores range from 1 ("not at all true of this lesson") to 5 ("very true of this lesson"). All models include fixed effects for raters, segments (or tasks), and randomization blocks, and cluster standard errors at the PST level. $p < .10$, p < .05, $p < .01$.

Table A7. Lee bounds with 95% confidence intervals

		Estimate	SE	Z	P-value	LL	UL
Performance Task SimSE	lower	0.95	0.14	7.00	0.00	0.68	1.22
	upper	1.00	0.13	7.53	0.00	0.74	1.26
Classroom SimSE	lower	0.62	0.17	3.61	0.00	0.28	0.96
	upper	0.85	0.17	5.02	0.00	0.52	1.19
Classroom QCI	lower	0.18	0.13	1.31	0.19	-0.09	0.44
	upper	0.31	0.14	2.17	0.03	0.03	0.58
Classroom MQI	lower	0.05	0.15	0.33	0.74	-0.24	0.33
	upper	0.26	0.14	1.79	0.07	-0.02	0.54

Note. Lee bounds with 95% confidence intervals. The bounds were estimated using the Stata leebounds command. Because the command does not allow users to include continuous

covariates, we estimated Lee bounds using the regression specification that excluded covariates but controlled for rater, segment (task), and randomization block fixed effects. Lee bounds should be compared to Appendix Table A5.

Appendix Table B1

	Reading and Reflection	Practice-Based
Section 1 Content: **Students with Mathematics Disabilities and Difficulties**	**Approximately 1.5 hours of asynchronous reading and written response:** Ball, K. (2020). First Person: Why my learning disabilities make me a better teacher. *Phi Delta Kappan*, *101*(5), 58-59. 1. How did teacher expectations, positive or negative, shape Ms. Ball's classroom experiences? 2. In the medical model of disability, the focus is on how an individual needs to be fixed. In the social model of disability, the focus is how society needs to be changed. Which model do you think Ms. Ball might support and why? Lewis, K. E., & Lynn, D. M. (2018). Against the Odds: Insights from a statistician with dyscalculia. *Education Sciences*, *8*(2), 63. 1. What did dyscalculia look like for Dylan? What types of activities or instructions were challenging for her? 2. Describe two ways that Dylan found to access ambitious mathematics instruction. 3. As a future elementary school teacher, what are some lessons that you can take away from these two articles for your own mathematics instruction? Center on the Developing Child at Harvard University. (n.d.). *A guide to executive function.* and National Center for Learning Disabilities. (n.d.). *Executive function fact sheet*. LD OnLine. 1. What are the three executive functioning skills? 2. List two challenges that students with developing executive functioning skills might face in a mathematics classroom 3. List two strategies a mathematics teacher might implement to support students with weaker executive functioning skills.	**Approximately 1 hour of asynchronous learning, including 17 interactive features and 3 interactive knowledge checks with immediate feedback:** · Personal account video (LeDerick Horne) followed by an interactive component (flashcards) describing possible noticings about students, teachers, and the learning environment. · Defining Disability: Video + Interactive knowledge check about social vs medical model of disability · Executive Functions and Mathematics Difficulties: Difference vs. Deficit graphic and video of Dr. Lewis, multimedia overview of working memory, cognitive flexibility, and inhibitory control · Interactive practice questions identifying examples of executive functions · Web comic about teacher perceptions, infographic about how teacher expectations impact mathematics learning, pictograph about racial disproportionality in special education · Overview of strategies for building on student strengths and interests, creating a supportive learning environment, and supporting students' needs · 8 item Section Quiz (100% accuracy required)

Section 2 Content: Word Problems	**Approximately 1.5 hours of asynchronous reading and written response:** Kennedy, M. J., & Romig, J. E. (2021). Cognitive Load Theory: An Applied Reintroduction for Special and General Educators. *TEACHING Exceptional Children.* Pages 1 and 2 only. 1. Briefly summarize Cognitive Load Theory. Stevens, E. A., & Powell, S. R. (2016). Focus on Inclusive Education: Unpacking Word Problems for Diverse Learners: A Guide to Using Schemas: Brad Witzel, Editor. *Childhood Education*, *92*(1), 86-91. 1. Describe the difference between a combine and change type of problem. Pongsakdi, N., Kajamies, A., Veermans, K., Lertola, K., Vauras, M., & Lehtinen, E. (2020). What makes mathematical word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills. *ZDM – Mathematics Education*, 52, 33–44. https://doi.org/10.1007/s11858-019-01118-9 1. Why might word problems be challenging for students to solve? Provide at least two characteristics of word problems that might impact the ease with which students can solve them **Synthesis Question:** How can schemas support students with word problems? Please include a definition of schemas.	**Approximately 1 hour of asynchronous learning, including 25 interactive features and 4 interactive knowledge checks with immediate feedback:** · Interactive overview of word problems with clickable grocery store examples · Multimedia overview of benefits and complexities of word problems, with classroom vignettes. · Explanatory video clips sourced from larger projects' expert panel · Differentiating between unpacking and solving word problems · Interactive exploration of word problem types/structures, with a focus on additive word problems (change and compare). · Multimedia overview of word problem features that impact complexity and cognitive load (ease of direct modeling, level of abstraction, language demands, quantities and units) · 8 item Section Quiz (100% accuracy required)

Section 3 Content: Metacognitive Modeling in Mathematics	**Approximately 1.5 hours of asynchronous reading and written response:** Quigley, A., Muijs, D., & Stringer, E. (2018). Metacognition and self-regulated learning: guidance report. Education Endowment Fund. Pages 6-17 only 1. What is the metacognitive regulation cycle? 2. What is modelling and how can it support the teaching of metacognitive regulation? Rhodes, S. (2019, December 12). How did you solve it? Metacognition in mathematics. ASCD. https://www.ascd.org/el/articles/how-did-you-solve-it-metacognition-in-mathematics 1. What is the role of metacognition in problem solving? Recommendation 2 Excerpt from Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A., ... & Ogbuehi, P. (2012). Improving Mathematical Problem Solving in Grades 4 through 8. IES Practice Guide. NCEE 2012-4055. *What Works Clearinghouse*. 1. Look back at the Quigley et al., (2018) article and the discussion of planning, monitoring, and evaluating. Which, if any, of these elements do you see in the teacher think aloud on page 20 (page four of your handout) of the Woodward article? Trocki, A., Taylor, C., Starling, T., Sztajn, P., & Heck, D. (2014). Launching a discourse-rich mathematics lesson. *Teaching Children Mathematics*, *21*(5), 276-281. 1. What do you see as the differences between Model Think Aloud 1 and Model Think Aloud 2? **Synthesis Questions:** 1. What are three tips you would give a teacher just starting out with metacognitive modeling for word problems?	**Approximately 1 hour of asynchronous learning, including 18 interactive features and 4 interactive knowledge checks with immediate feedback:** · Clear definitions of metacognition, metacognitive strategies, and the focal teacher practice of metacognitive modeling · Examples and non-examples followed by a sorting knowledge check · 4 video representations from a real classroom with decompositions (start by providing and then use knowledge checks for learner to decompose) · 2 more video representations from other teachers and grades with decomposition of practice · Guided example with knowledge checks · Interactive breakdown of steps for planning, tip sheet with sample language, and practice opportunities with example scripts provided. · 5 item Section Quiz (100% accuracy required)
Follow Up	**Module Reflection Prompts:** 1. What, if anything, was new to you in this content? 2. What, if anything, do you think you will take with you into your teaching? 3. What questions are you left with? Received Strong or Weak response feedback on all reflection questions	**Two Simulation Sessions with Coaching (2nd grade and 4th grade)** · Simulation Round 1 (~7 min) · Coaching (~10 min) · Simulation Round 2 (~7 min) · Feedforward

Coaching (~10 min)