Wisconsin Center for Education Research
SCHOOL OF EDUCATION
UNIVERSITY OF WISCONSIN–MADISON

Simulating Classroom Interactions at Scale for the Improvement of Practice-Based Teacher Education

WCER Working Paper No. 2022-3

December 2022

Courtney Bell, Geoffrey Phelps, Dan McCaffrey, Shuangshuang Liu, Barbara Weren, Nancy Glazer, and Francesca Forzani
Wisconsin Center for Education Research
University of Wisconsin–Madison
courtney.bell@wisc.edu

Key words: measurement of teaching, teaching quality, teacher education, assessment, teacher quality, performance assessment

Suggested citation: Bell, C., Phelps, G., McCaffrey, D., Liu, S., Weren, B., Glazer, N., & Forzani, F. (2022). *Simulating classroom interactions at scale for the improvement of practice-based teacher education* (WCER Working Paper No. 2022-3). University of Wisconsin–Madison, Wisconsin Center for Education Research.

Simulating Classroom Interactions at Scale for the Improvement of Practice-Based Teacher Education

Courtney Bell, Geoffrey Phelps, Dan McCaffrey, Shuangshuang Liu, Barbara Weren, Nancy Glazer, and Francesca Forzani

Abstract

The recent turn toward core practices and practice-based teacher education has been accompanied by a growing literature on the definitions, pedagogies to teach, and assessments of core practices. Despite these developments, the field lacks core practices performance assessments designed to be used across course sections, courses, and subjects. This paper provides an existence proof of this type of assessment and investigates the affordances and constraints of the approach. The study describes three types of mixed-reality simulation-based performance tasks of three core practices. More than 400 novices in 64 teacher preparation programs in the United States reported that they were able to use the simulation environment and believed the tasks measure important teaching skills. Scores on the tasks were positively related to novices' prior academic and teacher education experiences. Implications for the formative use of such simulations are discussed.

Simulating Classroom Interactions at Scale for the Improvement of Practice-Based Teacher Education

Courtney Bell, Geoffrey Phelps, Dan McCaffrey, Shuangshuang Liu, Barbara Weren, Nancy Glazer, and Francesca Forzani

Introduction

Within teacher education there is a “turn away from an intense focus on the knowledge needed for teaching to a focus on the use of that knowledge in practice” (Grossman, 2018). The goal of focusing on knowledge in practice is to help teachers learn “to enact teaching practices skillfully and knowledgeably” in order to “improve learning opportunities available for all students, and especially those from low-income backgrounds and minority groups” (p. 3).

Over the past decade, this turn—sometimes referred to as a focus on “core practices” or “practice-based teacher education”—has evolved. Since the publication of Grossman’s frequently cited paper (2009), which delineates the work of teaching from a cross-professional perspective, the literatures on core practices and practice-based teacher education have grown in complexity, to consider more subjects and a variety of teaching practices in both general and special education (Brownell et al, 2015). Those literatures now include science (Kloser, 2014; Stroupe et al., 2020; Windschitl et al., 2012), mathematics (Frank et al., 2007), history (Fogo, 2014), English/language arts (Grossman et al., 2013), and practices ranging from discussion to elicitation of student thinking, and classroom management (Ball & Forzani, 2011).

The turn toward core practices and practice-based teacher education is a curricular reform in teacher education. Its goal is to strengthen both what novice teachers learn and how they learn it. In short, the reform is designed to provide novices with reimagined opportunities to learn the teacher education curriculum. These reforms also strive to provide teacher educators with additional formative information about what novices know and can do. This information can then be used to adjust and improve a single course assignment, an entire course, or even a preparation program. For example, TeachingWorks, a group focused on supporting teachers and teacher educators, is partnering with 12 mathematics methods instructors around revising their methods courses and providing coaching to the methods instructors during the semester they teach the course (TeachingWorks, 2020). Beyond improving teacher education, some have argued that information generated from these reforms will lead to new understandings of teaching and teacher learning (Grossman, 2018) and provide novices with fresh insights during their training (see Janssen et al., 2014).

In the K–12 student realm, curriculum reforms frequently include (1) content, (2) pedagogies through which the content is taught, and (3) assessments that provide formative feedback about what students are learning. These categories overlap because the content one learns is always connected to how one learns it, and to the degree to which that learning is visible to others through assessments. In the teacher education realm, practice-based reforms also have these three dimensions of curricula. Core practices can be thought of as one aspect of teacher education content.[1]

Practice-based teacher education has pedagogies of enactment (such as rehearsals); and it is through those pedagogies that novices develop their ability to enact core practices. These pedagogies have evolved to include performance assessments that provide information to help teacher educators determine their next instructional steps. In a recent review of 18 avatar-based simulation studies published from 2013 to 2020, Bondie and colleagues (2021) found that all but four simulations were used in the context of a course and only one simulation did not provide coaching or feedback to the teacher.

The performance assessments used in the practices literature exist along a continuum. On one end, they are tightly tied to a specific context and on the other, they are more loosely tied to a specific context. Performance tasks developed and used within a single teacher educator's course would be on the "tightly tied" end. Hudson and colleagues (2019) used three increasingly challenging tasks in this way to better prepare special education teachers for managing the learning environment. Tasks developed and used across a program and/or teacher educators such as those being used at the University of Michigan in their elementary education program, or those at Oakland University in the secondary program, are on the other end of the continuum (Francis et al., 2018; Shaughnessy et al., 2019). At Oakland University, all secondary methods instructors assigned subject-specific weekly instructional tasks that provided formative feedback to novices across the program; the tasks were discussed as a whole program, biweekly, around common rubrics.

Despite these developments, the field lacks performance assessments designed to support insights about novice learning across courses, subjects, and certification areas within a teacher preparation institution (Cohen & Berlin, 2019). Such formative performance assessments can provide teacher educators with information to strengthen novices' opportunities to learn. That information could identify specific teaching practices or groups of novices needing continued support. Alternatively, novice reflections on and analyses of their performances on the assessments over time could be used by novices and programs to better understand how learning proceeds. This type of standardized, scalable performance assessment has many potential formative uses, but most uses will be focused at a level of aggregation that requires generalization—sections of a single course, course and field placement sequences, subjects, and certification areas.

This article has two descriptive goals: first, to provide an example of performance assessments that are designed for use across courses, sections of a single course, and subjects; second, to better understand the affordances and constraints of this approach. We do not believe simulations should replace other teacher education practices, but rather, seek to understand simulations so they might be used wisely toward the varied goals of teacher preparation

[1] Teacher education has many aspects of the curriculum or content—for example the development of novices' understanding of anti-racist teaching practices, child development, and schools as organizations. Core practices are taught alongside other content, not as a replacement for the other parts of the teacher education curriculum.

programs. To understand the affordances and constraints of the assessments, we consider novices' perceptions of the tasks, the degree to which tasks can be used at scale, and how task scores are related to other measures of novice teacher performance. We draw on both functional (Cronbach, 1988) and measurement validity perspectives (Kane & Wools, 2019) to describe the design of mathematics and English/language arts mixed-reality performance tasks focused on three core instructional practices—classroom discussion, eliciting student thinking, and modeling and explaining content (henceforth abbreviated DEME practices). We analyze novices' perceptions of these tasks as well as 12 performances from each of 414 novices in 64 teacher preparation programs across nine U.S. states, a larger and more diverse sample than previously documented in the literature (Bondie, et al., 2021). Finally, we consider how novice performances are associated with other aspects of novice preparation.

Approximating Core Practices via Simulation

The focus on core practices as a central feature of the curriculum of teacher preparation has led to numerous efforts to develop activities that approximate teaching. The general format for these approximations is often referred to as rehearsal, drawing attention to the potential to scaffold novices' learning opportunities through cycles of planning, enactment, and debriefing (Lampert & Graziani, 2009). Teacher educators across disciplines have begun to develop rehearsals, typically focused on the core practice of classroom discussion (e.g., Alston et al., 2018; Amador, 2017; Davis et al., 2017; Kavanagh et al., 2019; Kazemi et al., 2015; Kloser et al., 2019). While these rehearsals can vary, they often are designed around a "role-play" simulation in which a novice teacher plays the role of the K–12 teacher and other novice teachers play the role of K–12 students. The simulations often require substantial preparation before the actual rehearsal begins (Schutz et al., 2018). Preparation activities may include clarifying the focus of the interaction and even scripting the K–12 student responses. The enactment of the actual rehearsal is flexible. For example, the role play can be paused at various points to provide opportunities for participants to share insights and explore ideas (Averill et al., 2016; Campbell et al., 2020; Davis et al., 2017). The simulation can also be "rerun" to provide novices with opportunities to re-try aspects of the focal practices. We expect that the embodied cognition novices experience in simulation—the thinking that results from and shapes our body's movements—enriches novice learning, just as is does among younger K–12 learners (Nathan, 2022).

While approximations designed to provide opportunities to learn have long been used by teacher educators (Forzani, 2014; Zeichner, 2012), the advent of virtual and mixed-reality environments has enabled new approaches to simulating practice. For example, human-in-the loop technologies allow a novice to interact with virtual students who are voiced and animated by a digital puppeteer, referred to as an "interactor" (Dieker et al., 2014). Using the computer's audio and camera, the interactor responds in real-time to the novice as the simulation unfolds. The student avatars can be enacted with distinct personalities (e.g., shy, outgoing), behavior profiles (e.g., disruptive, compliant), identities, and learning profiles. Avatars can be assigned different racial characteristics, cultural backgrounds, genders, language backgrounds and ages, allowing the simulation to more closely approximate K–12 students.

Both role-play and mixed-reality approaches are also highly responsive to real-time judgements, bodily actions, and interactions that arise during the simulation. The focus and direction of the simulation can follow the opportunities that arise from the teacher's or simulated students' words and actions. In doing this, simulations can be aligned to the instructional needs of novices by focusing on particular content and learning outcomes, presenting K–12 student profiles that shape these interactions, and allowing teacher educators to provide feedback during or after the simulation (Bondie, et al., 2021). This ability of the simulations to responsively align to the instructional goals and learning needs of a particular teacher educator and class of novices, is the same characteristic that makes these types of approximations less useful for providing a learning experience that is interpretable across contexts. For example, a simulation in one teacher educator's course tailored to the learning needs of a novice who provides support to multilingual learners will have limited utility for other teacher educators in the same program who are focused on practices such as coordinating discussions in STEM or enacting culturally responsive instruction.

Standardizing Simulations of Core Practices

Individual courses are the building blocks for novice learning. Teacher educators frequently need to answer questions at the course level as well as more aggregated levels. Questions might include: How can course assignments be improved so that novices enact course concepts more proficiently in diverse field placements? To what degree do different sections of the elementary methods course provide similar learning opportunities? What are the strengths and areas of growth for our post-baccalaureate certificate versus our baccalaureate certificate program? These types of questions require teacher educators to have evidence that can be generalized over contexts. Such evidence requires assessments that are more standardized than an individual instructor's course assessment. Regrettably, standardization often suggests rote learning, trivial content, and misguided accountability. But there is nothing about standardization that requires these outcomes and consequences. For the formative use proposed here, standardization ensures that tasks that should be comparable are comparable; that novices have access to the task's evaluation criteria; that the claims made from the assessments reflect the construct's scope; and that the assessment is fair to all novices (see AERA/NCME/APA 2014 for additional standardization specifications). These features are especially important because they guard against key validity threats of performance assessments such as construct irrelevant variance, construct underrepresentation, generalizability, and comparability (Marion & Buckley, 2016)

Our review of teacher education simulations suggests that role-playing and mixed-reality simulations are differentially standardized along particular dimensions of performance assessment design: the focus of the simulation, the enactment of the teacher and student roles, and the success criteria by which novice performances are judged. The focus of the simulation—i.e., what the simulation is about, its components, objectives, and definition of a complete simulation—should be explicit. The focus should also be sufficiently comprehensive to cover the components of the core practice to guard against construct under representation. Simulations should be consistently enacted. Because the simulations are an interaction (among students, the novice, and the content), the interactors enacting K–12 students have a direct impact on both the content and challenge of the simulation experience. Consistent enactment can be judged by evaluating the consistency of the simulated students' actions toward and responses to novices. Third, the characteristics of a successful performance will vary but should be explicit. A successful performance at the beginning of a course might look different than one at the end. Well-specified standards of success provide the basis for providing consistent and interpretable feedback on the relevant dimensions of teaching across performances (Moss, 2011). To illustrate these features of standardization, we describe four simulation examples. They are selected to represent different core practices, designs for simulating the core practice, and approaches to standardizing the simulation content, performance enactment, and basis for evaluating the performance.

Four Simulation Examples

1. Simulated Encounters Focused on Differences in Power and Privilege

Developed at Vanderbilt University, Self and Stengel's (2020) SHIFT simulation cycles include a live-actor interaction that "simulates a situation that is common in teaching and that foregrounds identity, positionality, and systems of oppression in an attempt to make them more visible" (Self & Stengel, 2020, p. 3). Each SHIFT encounter is video recorded; encounters are analyzed through group discussion with simulation participants. Each SHIFT encounter is developed over iterations of specifying and revising the focus, content, and learning goals. A simulation enactment cycle has five steps. In the first step, the novice teacher reviews a specified critical incident and prepares for the interaction. Then there is a 10–12-minute simulated encounter with a live actor. The live actors receive training that includes information on what novice teachers should learn from the particular SHIFT encounter, how to start the simulation, the essential actions they must take with every participating candidate, anticipated responses from teacher candidates, and how to end the encounter. Live actors are expected to provide a consistent and comparable performance across each interaction. The final three steps of the SHIFT cycle involve analyzing the interaction with other novice teachers who each participated in the same encounter, reviewing a video of the interaction, and then discussing the simulation experience. The shared simulation experience is the basis for both collective consideration and individual reflection on the quality of the simulation performance. Participants are expected to come to a collective and individual interpretation of the meaning of the encounter and of the associated standards for judging both productive and less-productive teaching.

2. Argumentation-Focused Discussion

Mikeska & Howell (2020) developed mixed-reality simulations using TeachLivE^{TM} human-in-the loop technology (Dieker et al., 2014). The goal for the novice teacher is to lead a group of five K–12 student avatars in an argumentation-focused discussion where the students interact with each other's ideas, reconcile incompatible claims, and work toward consensus. Before engaging in the simulation, novice teachers review supporting materials that describe the lesson context, student learning goals, information about students' background, and information about a specific scientific investigation that students are conducting. These materials might also describe prior work students have done to generate a hypothesis, collect evidence, and make claims. Specific directions are also given for what novices need to do to successfully complete the simulation. In the 20-minute interaction, the novices engage the students in an argumentation-focused discussion. For each simulation, interactors are trained on five distinct student profiles that include specific claims, evidence-based reasoning, and anticipated responses for each student. Interactors were also trained to enact simulations where students respond to and build on each other's ideas (Mikeska et al., 2019). To help ensure consistency, interactors participate in practice interactions. Five dimensions were specified for evaluating novice teacher performance: "(a) attending to students' ideas, (b) facilitating a coherent and connected discussion, (c) encouraging student-to-student interactions, (d) developing students' conceptual understanding, and (e) engaging students in argumentation" (Mikeska & Howell, 2020, p. 10). Raters were trained to code these dimensions using a rubric and results from this coding were used to provide structured feedback to teacher educators and novices.

3. Eliciting and Understanding Student Thinking

Shaughnessy & Boerst (2018) developed a live-actor simulation designed to assess the knowledge and skills used in eliciting and understanding student thinking. Before beginning the simulation, the novice teacher is provided 10 minutes to review an example of student work and consider questions and other teaching moves they might make when interacting with the simulated student. A human interactor plays the role of the student and is trained on general response orientations (e.g., give the least amount of information that is responsive to the pre-service teacher's question) and specific responses to the mathematical questions that the novice asks. The interaction lasts up to 5 minutes and is ended at any point when the novice is satisfied that they understand the student's thinking. Raters are trained to score performances on four dimensions: "(a) eliciting the student's process, (b) probing the student's understanding of key mathematical ideas, (c) attending to the student's ideas, and (d) deploying other moves that support learning about student thinking" (Shaughnessy & Boerst, 2018, p. 45). Multiple raters were trained on the scoring checklist and performances are double scored with disagreements discussed and reviewed to reach consensus. The tasks were used formatively as a part of series of tasks in the elementary pre-service teacher education program at the University of Michigan.

4. Discussion to Establish Classroom Norms

As part of a study designed to examine the effects of coaching models, Cohen et al. (2020) designed a human-in-loop, mixed-reality simulation focused on facilitating classroom norm setting discussions. During each 5-minute simulation, novices were told to use effective redirections to address off-task student behaviors (e.g., humming, taking calls, singing). Effective redirection was defined as timely, succinct, and calm based on the Responsive Classroom guidelines for behavior management (Responsive Classrooms, 2014). Interactors were trained to ensure that the off-task behaviors conformed to a predefined list of task behaviors and were delivered consistently across each simulation. Human raters were trained to use an observation protocol to assess the timeliness of redirections, the proportion of specific redirections, the succinctness of redirections, and overall quality (i.e., effective use of calm, warm and supportive redirection). All performances were double coded by trained raters and scores were used to guide the coach's feedback in a teacher education course on classroom management.

Variation in Standardization

These examples—summarized in **Table 1—**illustrate different simulation designs that support forms of standardized administration. However, standardization is not an either-or proposition and the simulations can (and should) vary in their degree of standardization. Variation in standardization will vary by how the assessment will be used. For instance, the SHIFT encounter includes extensive pre-assignment reading to familiarize novices with the context preceding the live encounter and a detailed questionnaire used to help both the novice and the live actor prepare for the encounter. In contrast, the eliciting student thinking task used at the University of Michigan presents a single example of students' work immediately before the simulation begins with the expectation that the interaction will focus on this work sample and the immediate content (Shaughnessy & Boerst, 2018). The extent to which directions for the performance are specified and bounded shapes the degree to which the task directions, supporting information, and goals are similar across administrations.

Standardization of interactor behaviors also varies across the four simulations reviewed. Interactor behavior can be standardized through guidance on 1) how to respond to anticipated novice actions, 2) what to say or do to enact a behavioral profile or represent a particular understanding of the content, and 3) how to respond if the novice is taking the simulation in undesired directions. In the classroom norms simulation, interactors were provided detailed guidance on the specific off-task behaviors to deliver, and when to start and stop a behavior. Interactors were also monitored for fidelity and provided training feedback with the goal of ensuring "that each candidate had the same number of opportunities to respond to similar off-task behaviors" (Cohen et al., 2020, p. 7). In contrast, the SHIFT simulation allows for live actors to preview information on how novices understand the planned encounter and adjust their interaction behaviors based on this information. The SHIFT interaction, therefore, deliberately varies from novice to novice so as to be responsive to the novice's initial ideas.

The criteria against which novice performances are evaluated play a role in task standardization. Three of the four simulations included explicit rubric expectations for behaviors that represent a successful performance. Rubrics were used by trained raters to identify, count, and/or rate the quality of specific behaviors. Various techniques were used to ensure that rating criteria are applied consistently across raters and feedback can be interpreted in similar ways across performances (e.g., rater reconciliation conversations, interrater agreement metrics). The SHIFT simulation did not include rubric guidance for evaluating the performance. Instead, novices were expected to understand and judge the interaction through individual and collective consideration. This was done to cultivate novice careful reflection and broad understanding of equity oriented classroom interactions.

The simulations described here vary in their designs based partially on different intended uses. The SHIFT simulation is designed primarily for local use and is deliberately a teaching and learning activity. Consistent with its use, the simulations do not provide the novice with success criteria ahead of time. SHIFT simulations provide similar but not the same experiences to novices. This is done to facilitate reflective discussion among novices participating in simulations. There are no course grades or coaching interventions based on the performance; therefore, it is not important that the same scoring criteria be given to novices and used in a standardized fashion every time the simulation is performed.

In contrast, the other three simulations—while they include both live actor and mixed-reality simulations—have similar uses to one another, providing feedback to the novice in a course or program and helping teacher educators make judgements about novice performance levels. They all have explicit rubric-based success criteria, and each standardizes interactors as well as the task rating processes. It is unclear how similar or different tasks are within projects, and the management of simulations were a part of research project (in addition to providing feedback to teacher educators and novices).

Because the simulation is not evaluated in ways that are consistent across administrations, SHIFT has limited utility in comparing encounters across courses, programs, and other contexts. This is appropriate given its goals. The argument-focused discussions are designed to engage novice teachers in leading argument-focused discussions with the goal of providing detailed and actionable feedback to both teacher educators and novices on multiple dimensions of the performance. Finally, both the eliciting student thinking and classroom norm-setting discussions are designed to be administered and scored consistently. Because these tasks are shorter and have less involved scoring procedures, they are also feasible for use in comparing sizeable groups of students across course sections or even programs.

***Table 1*. Teaching Content, Enactment Approach, and Success Criteria Used in Selected Simulations of Core Teaching Practices**

Aspects of core teaching practices		1. Encounters focused on differences in power and privilege (SHIFT)	2. Argumentation-focused discussion (Argumentation)	3. Eliciting and understanding student thinking (Eliciting)	4. Discussion to establish classroom norms (Norms)	Modeling, Discussion, Eliciting
Teaching Content	Core practice(s)	Integrated practices in response to racially salient teacher-student interactions	Facilitating group discussion that involves argumentation	Eliciting and interpreting student thinking	Facilitating group discussion to establish classroom behavior norms	1. Modeling and explaining content. 2. Leading group discussion. 3. Eliciting and interpreting student thinking.
	Subject grade level	K–12 general education	K–6 science and mathematics	K–6 mathematics	K–12 general and special education	K–2 and 3–6 English language arts & mathematics
	Course embedded	Augment regular course content	Course activity	Course assessment	Course activity and assessment	Not connected to a course
Enactment	Interaction mode	Live students	Mixed-reality simulation specialists	Live students	Mixed-reality simulation specialists	Mixed-reality simulation specialists
	Standardization of interactor responses	Simulation specialists trained to follow interactor protocols.	Simulation specialists trained to follow interactor protocols.	Simulation specialists trained to follow interactor protocols.	Simulation specialists trained to follow interactor protocols.	Simulation specialists trained to follow interactor protocols.
	Interactor responses certified	NA	NA	NA	NA	Simulation specialists pass certification assessments.
	Standardized monitoring procedures	NA	NA	NA	Research team observed interactions	Specialist interactions monitored by 3rd party observers
Success Criteria	Judgement criteria	Collaboratively developed by novice participants	Criterion-based rubric	Criterion-based rubric	Criterion-based rubric	Criterion-based rubric
	Raters	Novice participants	3rd party	3rd party	3rd party	3rd party

Note. References for each simulation are 1. SHIFT (Self & Stengel, 2020), 2. Argumentation (Mikeska & Howell, 2020), 3. Eliciting (Shaughnessy & Boerst, 2018), 4. Norms (Cohen et al., 2020)

Understanding the Validity of the Mixed-Reality Simulation Tasks

Each of these four simulation features are connected to the validity of the tasks. There is no single way to evaluate the validity of an assessment. Validation is best thought of as a specification of a set of inferences and an analysis of the support for those inferences, given the specific purpose of the assessment (Kane, 2013).

Validity Considerations

Our purpose is to understand the strengths and limitations of mixed-reality simulations in teacher education, and therefore, we draw on both a functional (Cronbach, 1988) and measurement validity perspective (Kane & Wools, 2019). The functional perspective of validity concerns itself with the usefulness of the assessment. In this case, there should be evidence that when mixed-reality tasks are used across a course, courses, or programs, the tasks provide information that supports teacher educators' insights about course and program strengths and weaknesses. We focus on three dimensions of utility that are foundational: the definition of core practice being assessed; the grainsize of information provided by the tasks' scores; and novices' views of the tasks.

Some assessments of core practices focus on a single subject matter (e.g., Mikeska & Howell, 2020) or a narrowly delimited teaching practice such as redirecting off-task student behavior (e.g., Cohen et al., 2020). There are strengths in these approaches, e.g., higher reliability, close connection between coursework and the assessment. However, teacher educators need scores, and therefore, tasks, that can be generalized over novices and course(s) to understand what and how novices are engaging opportunities to learn at a programmatic level.

Closely related to the definition of the core practice is the grainsize at which scores are provided. If the definition of the core practice is subject specific or a narrowly defined teaching practice, scores will have a meaning at that level. By extension, scores at a larger grainsize or high level of aggregation will be less useful for understanding a novice's actions at a finer grainsize. For example, the total scores on a novice's secondary science licensure test provide less useful information to a teacher educator than the scale scores of the same assessment. The latter provides some insight into the specific areas of science strength and weakness that might be systematically addressed, while the former provides information on how much science the novice knows overall.

Finally, for assessment information to be useful, it must be based on an assessment that the novice perceives as reflecting the work of teaching and their current capabilities. When teachers (and administrators) do not see the teaching assessment as measuring the work of teaching or perceive it as not reflecting their capabilities, they distrust the assessment information. Then the assessment is less useful than it might otherwise be. A recent example of this is the perceptions of scores from value-added models (Pressley et al., 2018)

In addition to a functional perspective on validity, the mixed-reality simulations also should have evidence of measurement validity. We propose using task scores for local purposes, i.e., a single program, course, or courses, so that the functional perspective will be more important than the measurement perspective (Kane & Wools, 2019). However, both perspectives should be attended to.

A key design feature that supports higher-level claims about mixed-reality performances is that the task itself is standardized. Because the simulation asks the novice to interact with a simulation specialist who animates the mixed-reality student avatars, there are three dimensions of the task that require standardization: the task itself, raters, and the simulation specialists. To support the claims necessary, tasks need to be similar enough that they are generally interchangeable. For example, if two tasks are used to determine how well a group of novices carry out small group discussion and one task is very difficult and the other very easy, it will be hard to know how to interpret the novices' level of skill. On the easy task, the novices might look like they are making good learning progress and on the harder task they might look like they have not made as much progress. Thus, tasks must be designed to be comparable.

The mixed-reality simulations are rated by human raters and therefore, raters need to be able to agree on the score a performance is assigned. This means that rater consistency and rater accuracy should be considered. Further, there should be evidence that the simulation specialists provided a standardized interaction for the novices. This may take the form of information about their training, monitoring, and/or measures of their comparability.

Finally, from the measurement perspective, we hypothesize that novices who are more proficient on the simulations should be more proficient enacting the core practice when they are teaching in their own classrooms. Such predictive evidence is often hard to come by due to logistics and the cost of obtaining the data. Thus, concurrent validity evidence can provide some evidentiary support for the tasks' validity. The simulation tasks should be related in predictable ways to novices' experiences and skills.

Mixed-Reality Simulated Tasks

Together with colleagues at TeachingWorks and the virtual reality company Mursion, Inc., researchers and developers from the nonprofit educational research and assessment organization ETS created mixed-reality performance tasks of three core practices in K–6 English language arts (ELA) and mathematics. The tasks assess the practices of modeling and explaining content (Modeling), leading group discussion (Discussion), and eliciting and interpreting student thinking (Eliciting) (see **Table 1**, last column). The tasks are "in-the-moment" performance assessments because they are designed to be self-contained, stand-alone tasks that provide the novice with all the information they need to perform the task. The tasks are rated by trained and monitored human raters.

Task Description

The three task types—modeling, discussion, and eliciting—are each defined at a moderate grain size. Moderate refers both to the amount of time required to carry out the task and the nature of the teaching task. Novices completed performances of the core practices in 6–12 minutes, on average. This represents a time shorter than most classroom lessons but longer than the turn-taking exchanges that are common in classrooms. This grain size is shorter than

Mikeska and Howell's science argumentation tasks in science (2020) and longer than both the eliciting and the classroom management interactions studied by Shaughnessy and colleagues (2019) and Cohen et al. (2020), respectively. The tasks are of a moderate grainsize because they require coordination across subject matter, student understandings, and teaching pedagogies. For example, in a modeling task a novice would be asked to "explain how to solve a multidigit addition problem, 275 + 143, that requires regrouping," and do this by "showing the connection between the standard addition algorithm and the base 10 blocks" (Stickler & Sykes, 2016, p. 46). In the discussion task (**Appendix A**), the novice is asked to "lead a discussion aimed at helping students develop their ability to make and support inferences. The purpose of the discussion is to probe student thinking by having the students support their thinking with evidence from the text and by having them respond to each other's ideas." This coordination across subject matter, teaching, and students is visible in the dimensions along which the tasks are rated (**Table 2**).

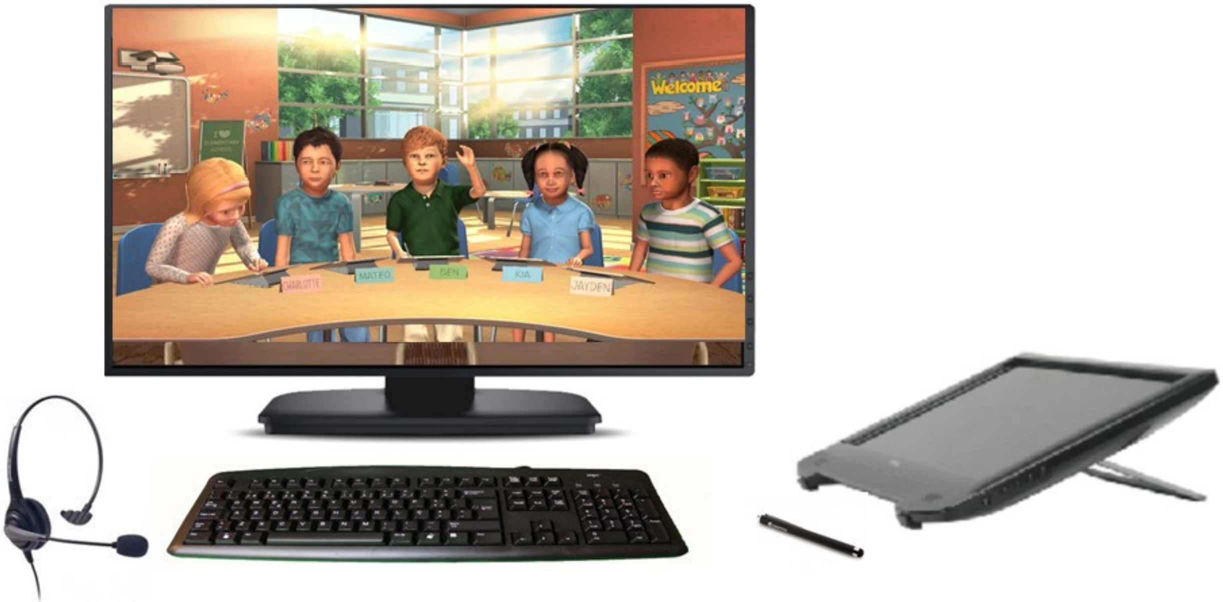
***Table 2*. Task Types and Scoring Dimensions**

Task Type	Scoring Dimensions
Modeling and Explaining Content	1. Framing the work
	2. Demonstrating the targeted process, strategy, or technique
	3. Narrating and annotating the demonstration of the process, strategy, or technique
	4. Using language, terminology, and representations
Leading Group Discussion	1. Eliciting and probing for each student's ideas
	2. Using students' ideas to steer the discussion toward the learning goals
	3. Representing content
	4. Summarizing and concluding the discussion
Eliciting and Interpreting Student Thinking	1. Using questions, prompts, and student tasks to elicit student thinking
	2. Attending to student talk and actions
	3. Interpreting student thinking
	4. Understanding the content

Because the goal of these tasks is to provide novice teachers with standardized opportunities to demonstrate their emerging competencies, the task materials and simulation specialist materials are tightly connected. The task materials define specific mathematics or ELA topics and specific research-based student ideas about those topics, which are presented deliberately by simulation specialists within task-specific guidelines. For example, in a task that requires the novice to lead a group discussion of a first-grade ELA text, the novice is asked to help students more deeply comprehend the text and "make inferences about Bindi's personality and support those inferences with evidence from the text." Prior to simulation, the novice reviews task materials that explain some text-based inferences about Bindi's personality (e.g., she is nice, creative, a good friend), what students might struggle with in making inferences (e.g., identifying evidence, connecting evidence with inferences), and requisite content knowledge for teaching (e.g., what a personality trait is, what an inference is). The associated simulation specialist materials provide general guidance, such as what to do in a situation when the novice asks a confusing question for that grade level, e.g., say “I don’t know” or ask a clarifying question. The simulation specialist materials also contain task-specific guidance. In the Kite Flight task, a common novice probing question to a student’s assertion that Bindi is “nice” might be “Where in the text do you have evidence that Bindi is nice?” Simulation materials included specific responses from the text (“Bindi asked Jack what was wrong”; “She stopped her bike when she saw Jack”) or the option to say “I’m not sure” or shrug the avatar’s shoulder. By tightly connecting the task’s teaching goal to background materials and simulation specialist materials, the simulation dimensions of the tasks are standardized to provide each novice a comparable opportunity to lead a group discussion.

To complete a task, the novice sits at a computer station that includes an electronic tablet, a keyboard, monitor, headphones, microphone, mouse, and a camera (see **Figure 1**). The simulation specialists, experts that are trained to interact through mixed-reality technology with novices in real-time, animate the avatars seen in Figure 1. Although simulation specialists interact in real-time with the novice, they are located remotely from the novice teacher. Remote may be in a room down the hall, across campus, or across the country. Some dimensions of the avatars are automated through technology (e.g., voice modulation, sitting postures, common movements such as raising a hand or writing). Other actions are created by the specialist moving in front of their own computer system, which is equipped with technology to detect the simulation specialist’s motions. Those motions are changed into avatar behavior and seen by the novices.

***Figure 1*. Novice Task Computer Station Set-Up**



For each task, the novice teacher is provided with information about her goal. For example, for a discussion task, the novice would be told the student learning goal and provided information about what happened immediately prior to the discussion the novice will lead. She would be told exactly what she was expected to do and the criteria by which her performance will be rated (see **Table 2** for scoring dimensions). Finally, the novice would be provided supplemental background materials about the subject matter in the discussion, common student challenges, and the relevant content knowledge for teaching. The novice would be given time to plan her approach and then begin the task. There is some variation in the specific preparation materials provided across the three task types, but all three take the same approach of clearly specifying what the novice should do, providing enough subject matter and student background information to do it, and allowing time for the novice to plan. Modeling tasks do not use avatars. Discussion tasks each had five student avatars. Eliciting tasks were carried out with a single student avatar. **Appendix A** includes a discussion task and its associated materials.

Task Development

Tasks were developed through an evidence-centered design approach (Mislevy et al., 2003). In this approach, the developer identifies a claim that scores should support; task design is oriented around eliciting evidence that supports the claim. The broad claim for these tasks was "Novice teachers who are more able to carry out the core teaching practice score higher on the assessment than novices who are less able to carry out the core teaching practice." To develop evidence for this claim, a multi-step process was engaged. To define the core practice construct, literature reviews for each core practice were conducted (Stickler & Sykes, 2016; Qi & Sykes, 2016; Witherspoon et al., 2016). These reviews linked teaching practices to valued student outcomes and identified how researchers have decomposed practices. Teaching is decomposed when we identify "essential elements of practitioner practice" (Resnick & Kazemi, 2019). Such decompositions are used for professional teaching and learning and by extension, for assessing progress in professional learning. In this case, insights from researchers' decompositions were the basis for task rubrics, thereby operationalizing quality criteria.

For each subject area, a national expert panel and survey of practicing teachers was also conducted. The panel and survey activities were designed to determine the ELA and mathematics content that practicing teachers viewed as both relevant and important to student learning (Martin-Raugh et al., 2016a; 2016b). Topics and student practices (e.g., reason abstractly, produce and distribute writing) were evaluated on the degree to which the content was foundational to the ideas and skills in the K–12 curriculum; taught in the targeted grades of K–6; occupied a large proportion of the curriculum; and fundamental to students learning in a way that if not taught well would be a source of difficulty in students' learning. There was also a national survey of practicing elementary teachers in ELA and mathematics to determine the degree to which certain core teaching practices were relevant and important to novice teachers' competencies (Martin-Raugh et al., 2016c).

Building on these construct development activities, tasks were developed collaboratively with TeachingWorks and Mursion. Tasks were made comparable by using task development guidelines that specified how to approach development for each task type. Tasks were piloted iteratively in five rounds over a 2-year period; each round piloted 2–8 tasks for each core practice. The goal of each pilot was to obtain roughly 20 participant performances and feedback regarding the communication technology, the avatars, and task materials for each task. Another goal was to gather simulation specialist performances and feedback. Developers used pilot performances and feedback to revise task and simulation specialist materials. They also created rating materials that could be used for main study rater training and simulation specialist monitoring materials for the main study. The discussion and eliciting tasks used avatars animated by simulation specialists as described above; however, the modeling tasks did not use an avatar.

Piloting resulted in both general and task-specific materials. The general materials used during the study were a warm-up attendance task, and training materials for raters and simulation specialists. These training materials taught raters and simulation specialists how to work with the online scoring system (raters), how to work with the mixed-reality technology (specialists), and the broad guidelines about interacting in any task (specialists).

Task Ratings, Training and Certification

The rating rubrics were designed to measure performance that could reasonably be expected of a beginning teacher. Tasks were rated on a three-point scale for each of the four dimensions of practice listed in **Table 1**. All three core practices were rated on similar rubrics (see **Appendix A** for one example). However, for the eliciting tasks, the novice completed the task and then completed three multiple-choice questions regarding what the virtual student knew about topic in the task. Raters used the multiple-choice responses (which were not rated correct or incorrect) together with the performance video to rate dimensions 3 and 4. No separate scores on the multiple-choice questions were created.

All performances were independently double scored. To create a total score for a performance, the average score for a dimension was calculated across raters and then added across dimensions. The highest score on a task was 12, the lowest 4.

Raters were trained, certified, and rated performances for one core practice online. Final training materials for each core practice contained three components: (1) project-specific training (including a rater guide and bias training), (2) practice materials on one math and one ELA task per core practice, and (3) task-specific materials on every task (eight for each core practice). Task-specific training materials included: the task, rubric and evidence inventory (rating notes), 3–4 annotated performances with rationales for the assigned scores, and a training set (a group of 2–4 performances given sequentially to raters in order to develop their understanding of the rating scales). After completing online self-training, raters had to pass a certification test for the core practice and were given two attempts to pass. Certification provided evidence the rater was competent to rate tasks for one type of core practice, however, given the specificity of each task, raters also were required to calibrate on individual tasks. Only raters who were trained, certified on the core practice, and calibrated to individual tasks were allowed to rate novice performances in the main study.

To be certified on a core practice, the rater rated two novice videos for one representative math task and two novice videos for one representative ELA task. To pass certification, raters had to have exact agreement with master ratings on at least half of the 16 ratings (4 dimensions for four performances), could have no discrepant ratings (>1 point difference from the master ratings), and had to have at least one exact agreement on every dimension of the task. Raters were also calibrated and allowed to rate specific tasks after passing a task-specific calibration.

The passing requirements were the same for calibration as for certification however, calibration was carried out on a single task. When raters were rating in the main study, they were monitored and supported by expert scoring leaders.

Simulation Specialist Interaction Standardization

To ensure that novices had similar opportunities to demonstrate their ability to carry out the discussion or eliciting practices, simulation specialists were trained, certified, and monitored. Because each simulation specialist needed to be sufficiently prepared to perform interactions for *each* task in their assigned practice, specialists were trained, certified, and monitored on each task. This resulted in the development of 16 sets of training and certification materials (eight discussion and eight eliciting). In order to be certified to interact on a specific task, the specialist had to pass a certification test in which the specialist acted out the avatars as they would in a normal task. A standardized novice, portrayed by a task developer, was used for the certification test to provide the simulation specialist with a fair (and similar) chance to certify. Simulation specialist performances were rated by trained raters.

Methods

The study design and analysis were carried out to address the three research questions that concerned novices' experiences with the mixed-reality tasks, the quality of task administration and ratings, and the relationship between novice task performances and other measures of novice teaching skills and experiences.

Study Design

The 24 tasks were administered to volunteer novice teachers over the course of 13 weeks and included 64 educational preparation programs in nine states at 26 computer labs. Four hundred fifty-five novice teachers completed two different six-task assessment sessions at two time points that were separated by no more than 2 weeks. To minimize concerns about novice learning across occasions, this analysis draws only on data from the first six tasks each novice completed. There are no noteworthy differences in findings if data from both occasions are used.

For six tasks (two of each core practice), novices were given up to 5 hours to complete the tasks in a secure testing center. **Table 3** shows the preparation time allowed for each type of task and the time taken by novices. On average, novices used 7.0 minutes for eliciting tasks, 11.3 minutes for discussion tasks, and 6.3 minutes for modeling tasks. Almost all novices provided basic background information on their race and gender (99%) prior to the first tasks, and almost all (96%) completed an online questionnaire regarding their testing experience after the assessment was complete. Novices were compensated for their time.

***Table 3*. Preparation and Performance Time for Core Practice Tasks**

Core Practice	Preparation Time	Mean Performance Time	Std Dev Performance Time
Modeling	20	6.3	2.4
Discussion	30	11.4	3.2
Eliciting	10	7	2.5

Sample

Three groups of elementary general education teachers were recruited to participate in the study: teachers who recently completed a teacher preparation program and were teaching elementary grades less than 1 year as the teacher of record, teachers who were about to graduate from their elementary preparation program but had not yet become a teacher of record, and teachers who were still in their elementary preparation program and had experience practice teaching (e.g., student teaching, substitute teaching, practicum teaching). Study participants resided in nine states (AR, CT, MD, TX, MI, NJ, MO, PA, SD).

The sample can generally be described as mostly White, female novices who were either in their final year of undergraduate education or completed at least a bachelor’s degree. Just under a quarter of the sample were teachers of color. More than 80% of novices were in their student teaching placements or had already completed at least one student teaching placement at the time they took the assessment. **Table 4** describes additional details.

The original sample contained 455 novices. After dropping those who did not complete the first six tasks (*n*=23) and those who did not have either background questionnaire or the assessment experience questionnaire (*n*=18), there were 414 novices in the analytic sample. All novices had complete data for the analyses.

***Table 4*. Gender, Race/Ethnicity and Preparation Pathway of Study Participants, Proportion of Sample (*N*=414)**

	Study Participants
Gender	
Female	91
Male	9
Race/Ethnicity	
American Indian or Alaska Native	1
Asian or Pacific Islander	3
Black or African American	12
Hispanic/Latino	3
White	76
Two or more races	5
Education completed	
Sophomore (2nd) year	1

Junior (3rd) year	6
Senior (4th or final) year	50
B.A. degree	20
B.A. + additional credits	16
Master's degree	5
Master's + additional credits	2
Student teaching	
Have not completed student teaching	18
Currently completed or have completed one or more student teaching experiences	82

Assessment Task Administration, Simulation, and Rating

Each novice performed six tasks. For each core practice, there was an upper and lower grades task; one task was ELA, the other mathematics. This means that each novice performed three upper and three lower grades tasks, as well as three ELA and mathematics tasks. Tasks were spiraled to account for task ordering effects. Analyses aggregate over all tasks and novices.

For the eliciting and discussion tasks, simulation specialists were trained and certified on 16 discussion and eliciting tasks; 94 out of 168 interactor/task pairs (56%) were certified on their first attempt; 52 interactor/task pairs (31%) were conditionally certified; and 22 interactor/task pairs (13%) did not pass. Conditionally certified simulation specialists and those that did not pass were given remediation. If they did not successfully certify on a second attempt, they were not allowed to interact for that task.

Eighteen simulation specialists were certified prior to the study. To ensure each novice had a similar opportunity to perform the core practice, the standardization of interactions was monitored during the 13-week pilot timeframe. Monitoring analyses suggest the specific simulation specialist a novice was paired with did not impact the rating the novice received.

Raters were trained and certified to ensure they were accurately using the rating scales. Certification rates varied by core practice: 57% for eliciting tasks, 70% for discussion tasks, and 75% for modeling tasks. Once certified, raters calibrated before each rating each day. Raters using observation tools have been shown to drift during main study rating (Casabianca et al., 2014) and less expert raters also tend to use criteria not in the rating scale (Bell et al., 2014). Calibration activities were used to mitigate these concerns. Calibration pass rates varied by core practice. If a rater did not successfully calibrate, they were not allowed to rate that type of task that day but could try again on the next day. Calibration pass rates were 64% for eliciting tasks, 88% for discussion tasks, and 71% for modeling tasks.

Novice performances were rated over a 15-day rating period by raters certified (by core practice) and calibrated (by task). They were also monitored and supported by 23 scoring leaders (six leaders for eliciting, 10 for discussion, seven for modeling). There was a scoring leader to rater ratio of 1:5. At the beginning of each rating shift (4, 6, or 8 hours), raters and scoring leaders refreshed their training by watching videos that had pre-scored for the task they would be rating that shift.

All task performances were double rated. A task's score on each dimension was calculated by averaging the dimension/rater ratings from two raters. Raters' notes and ratings were recorded in an online rating platform. Assignment of rater to performance (and novice) used a balanced rating design that carefully assigned raters to tasks and novices to maximize the number of different raters each novice's performances received.

For simulation tasks to provide information teacher educators can use for strengthening novice learning opportunities, it is important that raters use the full scales in reliable ways. To determine whether raters could use the scales consistently, the proportion of rater agreement at the scale level for each task was calculated. Exact agreement implies two raters assigned the performance the same rating. Adjacent and discrepant agreement implies two raters assigned ratings one off from one another (adjacent) or two off from one (discrepant). Intraclass correlation coefficients were also calculated as a measure of rater reliability.

It is important for tasks to capture variation in performances. Scores that reflect the full range of the rating scale suggest that raters can detect differences in performances that have been theorized to exist during the task and rubric development phases of task design. Researchers have documented the lack of variation in one of the field's common assessment tools—observational rubrics (see BMGF, 2012; Liu et al., 2019) and raised concerns about what this means for improvement efforts (TNTP, 2008). Descriptive statistics were calculated for each task to determine the degree to which raters used the entire rating scale.

Novice Questionnaires

Novices completed two questionnaires that gathered background information and perceptions of the experience and authenticity of interacting with the avatars and performance tasks. The first questionnaire was brief (~15 minutes) and focused on demographics (e.g., gender, education, race/ethnicity), certification area, preparation program description and teaching of high-leverage practices, teaching experience, and student teaching experience (where applicable). The first questionnaire was administered when the novice signed up to participate in the study. The response rate was 99%.

The second questionnaire was administered after completion of the second assessment session and required an average of 10 minutes to complete. It focused on the teacher's views of the three types of core practices tasks, the appropriateness of the avatar behavior, the clarity of the tasks, and the similarity of the tasks to real tasks of teaching. The response rate for the second questionnaire was 96%.

To determine novices' perceptions of the mixed-reality performance tasks (research question 1), responses from the post-assessment questionnaire were aggregated across novices and reported on the Likert-type scale of the questionnaire.

Other Measures of Novice Knowledge and Experience

Student teaching plays an important role in novices having the opportunity to enact core practices. On the post-assessment questionnaire, novices reported on whether they were engaged in student teaching (or had already completed it) or they had not yet carried out student teaching. Many of the novices were in the late spring of their programs and were therefore engaged in student teaching. Novices also learn about subject matter, and in the case of teacher education students, teaching practices during their undergraduate and graduate educations. Novices also reported on their grade point averages (GPA) on the post-assessment questionnaire.

To facilitate analyses of the relationships between novices' performances and other measures of their knowledge and experience, total scores for the core practices tasks were created by summing across average dimension ratings for all six tasks. The highest total score on six items was 72. To specify the association between total core practice scores and student teaching or GPA, Pearson correlations were calculated. ANOVAs and multiple pairwise t-tests were carried out to determine whether associations were statistically significant.

Findings

What Were Novices' Perceptions of the Simulation Tasks?

These in-the-moment mixed-reality simulations were designed to allow novices to enact core practices important to student development. For novices to demonstrate their ability to enact such practices, they should understand the task, have enough time to prepare for and carry out the task, and feel able to use the mixed-reality environment competently. Task validity is further enhanced if novices feel the task measures something important. On a 4-point Likert-type scale (*strongly disagree*, *disagree*, *agree*, and *strongly agree*), novices were asked about the degree to which they agreed with statements in **Table 5**. Due to the pattern of responses, novices' views were collapsed into two more conservative categories, *agree* and *disagree*.

Almost all novices (97%) reported that the assessment materials were clear. Most novices (89%) found the assessment interface was easy to use and reported feeling comfortable interacting with the avatars. At least nine of 10 novices also felt that the tasks were authentic and measured important teaching practices. Slightly more than three-quarters of novices felt that their performances on the tasks accurately reflected their actual capability in enacting the three practices. It is unclear why almost a quarter of novices did not feel their performance accurately reflected their capability. This is especially interesting because they reported giving their best effort and the tasks were clear, the interface easy to use, and avatar behaviors were typical of student behaviors. This could be because the novices felt they could have performed better for reasons that have little to do with the assessment and more to do with their own personal expectations of themselves. Alternatively, perhaps those novices who found the tasks somewhat unclear and the assessment interface difficult to use felt they performed less well than they are capable of in other settings. The current data do not allow us to understand novices' views of their performances in more nuanced ways, however, their views are important to better understand.

***Table 5*. Candidates' Views of Tasks and Simulation Experiences (*N*=414)**

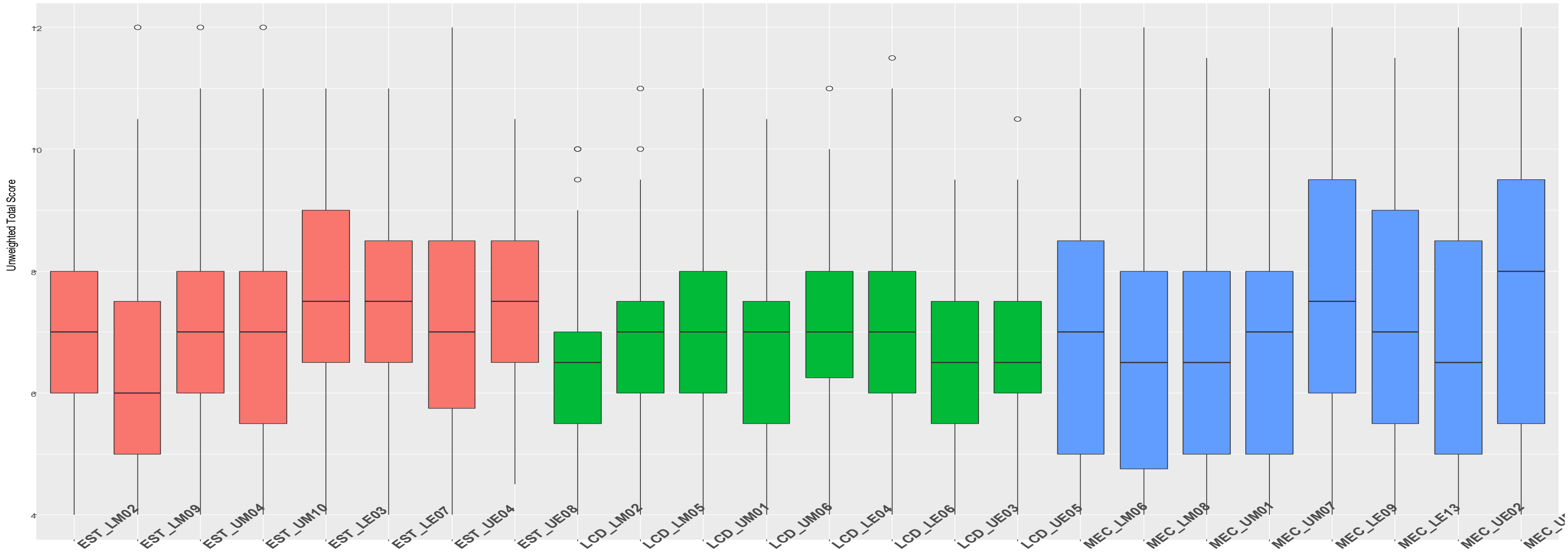
Question	Agree	Disagree
The materials provided for use during preparation for the tasks (e.g., task directions, scenarios, and goals) were clear.	96.9	3.1
I found the overall testing interface, such as the touch screen and writing tools, easy to use (even if the tasks themselves might have been difficult).	88.6	11.4
I felt comfortable interacting with the simulated student(s) on the computer.	86.2	13.8
I found the touch screen's shared work-space and writing tools easy to use with the simulated student(s).	91.8	8.2
The kinds of teaching skills or abilities required by the tasks felt authentic to me.	85.7	14.3
The tasks assess important teaching skills or abilities.	91.1	8.9
The simulated students' responses and behaviors during the tasks were typical for students at their grade level.	94.9	5.1
For the Modeling and Explaining Content tasks (no simulated classroom), the touch screen was an acceptable alternative to a whiteboard and did not significantly affect or change how I would teach the lesson.	89.4	10.6
For the Eliciting Student Thinking tasks (with one simulated student), the Post Performance Questions gave me a good opportunity to show what I had learned about the student's thinking.	91.3	8.7
Some of the tasks were difficult for me to complete successfully because I was unfamiliar with the technology.	31.4	68.6
Some of the tasks were difficult for me to complete successfully because I was unfamiliar with the content topic.	33.8	66.2
My performance on the tasks accurately reflects my ability to employ these teaching practices.	76.3	23.7
I gave the tasks my best effort.	99.8	0.2

To What Degree Were the Tasks Able to be Administered and Rated in Valid Ways?

Variation in Scores

Raters assigned ratings that generally used the full scoring scale. **Figure 2** summarizes the descriptive statistics for all tasks in box plots (the associated table for the box plots is in **Appendix B**). Each task is a different box plot. Eliciting tasks are shown in red, Discussion tasks in green, and Modeling in blue. The box for each task shows the middle 50% of scores, with the whiskers showing the lower and upper 25%. The small circles at the extremes of the scoring scale show outliers. Most of the tasks piloted showed variation in scores with novices earning scores from the lowest to the highest score. For eight of 24 tasks, no novice reached the highest score. The discussion tasks had a somewhat restricted range, with some of the highest scores achieved infrequently or not at all.

***Figure 2*. Task Scores for All Piloted Tasks (24 tasks, N=414)**



The competency demonstrated in performances varied by task type, subject, and grade span (**Table 6**). In general, discussion tasks had lower mean scores as compared to the other two core practice tasks, suggesting that the group discussion tasks were somewhat more difficult for novices. Mathematics tasks were more difficult than reading/language arts tasks. In pairwise t-tests, these two differences were statistically significant at the 0.05 level. The upper grades tasks had lower mean performance scores than the lower grades tasks, however, the differences were not statistically significant or of notable size.

These results may reflect something specific about the tasks themselves; for example, all the ELA tasks may have unintentionally been written in such a way that they were easier for novices. It is also possible that novices are simply more skilled with the practices when they occur in ELA. If novices were more comfortable with the ELA subject matter they might be better able to show competency with the teaching practices in the ELA context. Alternative explanations for both the subject matter and task type differences can be developed (e.g., perhaps discussion is a more challenging core practice or working with a single student—as in the eliciting tasks—is easier than working with a group of students). The current data do not allow us to sort out these potential explanations, but it is important to note that there were differences in scores by two features of teaching: subject matter and core teaching practice.

***Table 6*. Scores by Task Type, Subject, Grade Level**

***N*=414**	**Min**	**Max**	**Mean**	**Median**	**SD**
Task Type					
Eliciting	8.0	22.5	14.2	14.5	2.7
Discussion	8.0	20.5	13.6	13.5	2.5
Modeling	8.0	24.0	14.1	14.0	3.5
Subject					
Reading/language arts	12.0	34.0	21.6	22.0	3.9
Math	12.0	31.0	20.2	20.0	3.7
Level					
Lower	12.0	31.0	21.1	21.0	3.8
Upper	12.0	31.0	20.8	20.8	3.8

Reliability

It is important for the tasks to be able to consistently measure novices' capabilities. Depending on the specific task, subject, and dimension of performance being rated, exact agreement on dimensions ranged from 58 to 70% (**Table 7**). Raters rarely disagreed by two points and in general, raters assigned dimension ratings that were within one score point of each other 95–100% of the time. These moderate levels of exact agreement are similar to or slightly better than observation-rubric studies of naturally occurring classroom teaching (BMGF, 2012).

***Table 7.* Agreement Rates for Tasks Across Dimensions**

Core Practice	Dimension	Exact	Adjacent	Discrepant	Exact + Adjacent
Eliciting	Dimension1	59	39	2	99
	Dimension2	60	39	2	99
	Dimension3	63	36	1	99
	Dimension4	61	38	2	99
Discussion	Dimension1	64	36	1	100
	Dimension2	58	40	2	98
	Dimension3	59	40	2	98
	Dimension4	70	30	1	99
Modeling	Dimension1	65	33	3	98
	Dimension2	63	35	3	98
	Dimension3	61	37	3	97
	Dimension4	58	40	3	98

How Did Scores on These Tasks Relate to Other Measures of Novice Performance?

If the tasks are measuring teaching practices in valid ways, we would expect that novices' scores would be related to novices' other relevant knowledge and experiences. Therefore, we investigate the relationship between novices' GPA, their student teaching experience, and their performance scores. We would expect that novices who have been more successful in their undergraduate programs may have higher levels of knowledge about students, subject matter, and/or teaching practices than novices who had less success in their undergraduate course experiences. More successful undergraduates may also have developed higher levels of teaching competency in their various practice-based teaching placements, prior to student teaching. Thus, we predict that higher GPAs will be associated with higher novices scores across the six tasks. This is what we find (**Table 8**).

On average, novices with higher GPAs tended to have higher task scores. A multiple group comparison using ANOVA suggests that there were significant differences in total test scores across the three GPA groups ($F=10.98$, $p<0.0001$). As the ANOVA test shows significance across groups, we computed Tukey Honest Significant Differences to run multiple pairwise-comparisons between the group means. We find that the 3.5–4.0 GPA group performs significantly better than the two other groups, the 3.0–3.49 GPA group does not have a significantly higher average score than the 2.0–2.99 group. This latter finding may be due to the small number of novices in the 2.0–2.99 GPA group.

***Table 8*. Novices' GPAs and Mean Total Scores (*N*=414)**

GPA	n	Mean Total Score	Std Dev
3.5–4.0	270	43.12	6.75
3.0–3.49	126	39.73	6.16
2.0–2.99	18	38.11	7.38

We might also expect that having had opportunities to work with K–12 students will result in greater competency with the core practices. Many programs incorporate field experiences into novices' courses from the very first course of the preparation program (Hollins, 2015). However, the student teaching experience remains a critical learning experience for novices because for many, student teaching is their first opportunity to repeatedly plan and implement full lessons and (sometimes) units of instruction with an entire classroom of students. We predict that novices who are currently engaged in or have already completed their student teaching experience will have had more opportunities to practice the three core practices and therefore will earn somewhat higher scores on the tasks. **Table 9** shows that this prediction is supported empirically. While there are only 64 novices who have not engaged in student teaching, their mean total scores on the assessment are more than two points lower than the average novice's score who has had the opportunity to engage in student teaching. This difference is roughly one-third of a standard deviation. A two group t-test shows that this difference is statistically significant at the .05 level (t = 3.26, p-value = 0.0015).

***Table 9*. Novices' Student Teaching Experience and Mean Total Scores (*N*=414)**

Student Teaching	n	Mean Total Score	Std Dev
Yes	341	42.27	6.59
No	73	39.99	7.56

Discussion

This study found that the mixed-reality simulation tasks trialed in this study were positively regarded by novices, were able to be enacted and rated in standardized ways and were related to novices' previous undergraduate and teaching experiences in predicted ways. Specifically, novice teachers were able to use the mixed-reality simulation technology and generally felt able to understand what was asked of them. They also reported that the simulation tasks were authentic and captured teaching practices that they view as important to teaching and learning. The enactment of the two core practices that used simulation specialists—discussion and eliciting practices—were able to be rated reasonably reliably by raters. Rater agreement and interrater reliability was similar to or better than large-scale observation scoring projects (see

BMGF, 2012). Finally, novices' GPA and student teaching experience were positively related to their scores on the simulation tasks.

This study contributes a first-of-its-kind demonstration of the standardized assessment of core practices at scale. However, as the practice-based teacher education field evolves it is important to consider both the affordances and constraints of these types of tasks. We now turn to a discussion of affordances and constraints, taking care to address the perspectives of functional and measurement validity.

Each of these tasks, which includes task, rating, and simulation specialist materials, had clearly articulated standards of quality and as administered in the mixed-reality environment, resulted in reasonable levels of interrater agreement and consistency. This suggests the tasks have some foundational validity evidence that would allow them to be used formatively to understand what students are learning in a single course, across sections of the same course or even a program. This is an affordance of the tasks because the field does not yet have such empirically studied performance tasks, but it is also a potential constraint.

The tasks require considerable training and administrative effort to coordinate, and they are time consuming for candidates when preparation, performance, and questionnaire time is considered. If teacher educators used the performance to formatively improve novices' learning opportunities by discussing patterns with their colleagues, this would add additional time and effort for pattern analysis and discussion of next steps. Assessment time in any preparation program is limited and performance tasks should be undertaken with care. While these tasks are time consuming (and likely more so if used formatively), they may be less time consuming than some of the portfolio assessments currently used to provide program feedback (Bergstrand et al., 2017).

An additional affordance and constraint of these tasks concerns the mixed-reality environment. The tasks are interactive and thereby responsive to the novice; avatars can be changed to provide a wide range of learning opportunities to novices. Further, the mixed-reality environment allows many people to fill the role of the simulation specialist. However, carrying out these tasks in a standardized mixed-reality environment is expensive. There are both platform and technology costs because the intellectual property that makes the mixed-reality environment possible is not in the public domain. As previously mentioned, in teacher education programs around the country, live-actor role-play simulations are already being used by teacher educators and are an alternative to the mixed-reality environment. It is unclear how the affordances of the mixed-reality environment should be compared to the affordances of a live-actor simulation. How would time, expense, and task flexibility change if live-actor role-play simulations rather than mixed-reality simulations were used at this level of standardization? If certain aspects of rating, and rater/simulation specialist monitoring were simplified, it is possible a live-actor mode would be less burdensome than the mixed-reality mode. There may be tradeoffs between utility (i.e., administering the tasks with simplified approaches and live actors) and reliability and validity (e.g., rater agreement, relationships to novices' experiences as undergraduates and as student teachers). These trade-offs would need to be considered and if possible, remediated.

A final affordance concerns the potential contributions to scholarly knowledge of these tasks and performance assessments like them. Researchers have claimed that practice-based teacher education can help us learn about teaching, teacher learning, and teacher education (e.g., Hiebert & Morris, 2012). If the results of this study generalize over other studies and core practices, these particular standardized tasks allowed us to notice three insights about the nature of teaching and learning to teach. Those insights are that: 1) novices performed worse on mathematics tasks than they did on ELA tasks, regardless of grade level or core practice; 2) novices performed similarly in the lower grades tasks and upper grades tasks; 3) novices performed better on modeling and eliciting practices as compared with discussion practices.

Each of these insights has multiple plausible explanations. For example, task type variability could explain why discussion tasks scored the lowest of the three tasks; discussion tasks were differentially hard due to task design. Alternatively, higher performance on eliciting and modeling might be a result of novices having more opportunities to learn these practices and fewer opportunities to learn the discussion practice. Alternatively, it may also be explained by the order in which novices learn core practices. Perhaps discussion requires some of the skills embedded in modeling and eliciting and therefore, takes longer to learn, sequentially reaching a proficient level after the other tasks.

Each of these teaching insights is facilitated by standardized tasks and clear specifications of quality. By having standardized tasks, teacher educators can know what novices can do at a given moment in time and map those capabilities to students' opportunities to learn in a course, courses or program. While practice-based research has existed for more than 10 years, the field still needs tools that allow us to do just this type of discernment at levels of aggregation above a single teacher educator's section of a course. The findings from this study, which suggest that both subject matter and practices themselves may provide varying difficulty for novice learning, suggest there is much to be learned. Such insights might nominate ways to accelerate novice learning. If true, this would be especially important to know, given the number of novices in their first years of teaching who disproportionately serve our most academically disadvantaged students.

Study Limitations

Like every study, there are limitations of these findings and what we can learn from the study. This paper reports on the tasks themselves and a specific administration of those tasks. The tasks may or may not be interpreted by all novices in similar ways or exhibit the distributions and relationships documented here under different administration, training, and monitoring conditions. To understand the robustness of these findings, the tasks should be tested with other populations and circumstances. Further, the tasks were not used to provide formative feedback to a teacher educator(s) about a specific group of their students. If the tasks are to be useful for the formative purpose envisioned, they will need to be studied under those conditions and a persuasive validity argument should be developed (Kane, 2013). Both these caveats suggest that the evidence provided here is promising but should not be interpreted to generalize to contexts beyond this study.

While the evidence provided here represents nearly 5 years of development, fielding, and research, it is still nascent. As noted in the discussion, additional research is necessary to understand the full potential of these types of tasks. In particular, stronger, and preferably quasi-experimental or experimental evidence is needed to support the relationship between novices' performance of these core practices and their eventual performance in classrooms.

Conclusions

The tasks described here and the large-scale data collected provide an existence proof of one way to address the field's need for formative assessments that can be used to understand within and across course learning by novices. This study documents how novices engaged in and perceived three core practices and looks carefully at subject matter, grade band, and core practice. We view these findings as preliminary, but promising. And they should be seen as a part of a larger body of research investigating the utility of simulations in higher education (Chernikova et al., 2020).

We are, however, cognizant of the urgency we face in U.S. teacher education. While we agree with Hiebert and Morris (2012) that practice-based teacher education should be situated in local instructional contexts that accumulate and pass on records of practice (they argue for annotated lesson plans and assessments), we take Zeichner's (2012) point too. He argues that a practice-based approach to the teacher education curriculum is necessary, but not sufficient to address the longstanding inequalities of U.S. public schools. Our students and novices urgently need us to work together and at-scale to strengthen novices' opportunities to learn so they are prepared to address the injustices our current school systems perpetuate. For the community of practice-based teacher educators working daily to build practice-based curricular and pedagogical innovations (see Francis et al.; Grossman, 2018; Self & Stengel, 2020), we offer up the assessment approach of these tasks as one way to further deepen and extend their innovations.

We recognize this will take time. The years required, and the dozens of professionals that worked on these 24 tasks, underscores the importance of time and significant human capital required to develop rich, practice-based, scalable teacher education assessment tools. We have discussed some ways that it might be possible to simplify this approach, but even with such simplification, it may be many years before the field is able to develop the full range of formative tools we need to understand within and across course practice-based opportunities to learn. We look forward to participating in the community of teacher educators and researchers engaging in that work.

References

Allen, D. (1967, September). Microteaching, a description. ERIC document (ED 019 224). http://files.eric.ed.gov/fulltext/ED019224.pdf

Bondie, R., Mancenido, Z., & Dede, C. (2021) Interaction principles for digital puppeteering to promote teacher learning, *Journal of Research on Technology in Education, 53*(1), 107–123. https://doi.org/10.1080/15391523.2020.1823284

Ball, D. L., & 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

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Gitomer, D. H., McCaffrey, D. F., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 50–97). Jossey-Bass.

Bergstrand Othman, L., Robinson, R., & Molfenter, N. F. (2017). Emerging issues with consequential use of the edTPA: Overall and through a special education lens. *Teacher Education and Special Education, 40*(4), 269–277. https://doi.org/10.1177/0888406417718251

Bill & Melinda Gates Foundation (BMGF). (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains.* Author.

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

Clifford, H., & Edwards, C. H. (1975). Changing teacher behavior through self-instruction and supervised microteaching in a competency-based program. *The Journal of Educational Research, 68*(6), 219–222. http://search.proquest.com/eric/docview/1290444730/fulltextPDF/9C5435F17A154237PQ/1?accountid=14739

Cohen, J., & Berlin, R. (2019). What constitutes an "opportunity to learn" in teacher preparation? *Journal of Teacher Education, 71*(4), 434–448.

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%2F0162373720906217

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Erlbaum.

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

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. https://doi.org/10.1177/0888406413512683

Fogo, B. (2014). Core practices for teaching history: The results of a Delphi Panel Survey. *Theory & Research in Social Education, 42*(2), 151–196. https://doi.org/10.1080/00933104.2014.902781

Francis, A. T., Olson, M. R., Weinberg, P. J., & Stearns-Pfeiffer, A. (2018). Not just for novices: The programmatic impact of practice-based teacher education. *Action in Teacher Education*, *40* (2), 119–132. https://doi.org/10.1080/01626620.2018.1424053

Ghousseini, H., & Herbst, P. (2016). Pedagogies of practice and opportunities to learn about classroom mathematics discussions. *Journal of Math Teacher Education, 19*, 79–103. https://doi.org/10.1007/s10857-014-9296-1

Grossman, P. (2005). Research on pedagogical approaches in teacher education. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying teacher education* (pp. 425–476). American Educational Research Association.

Grossman, P. (Ed.) (2018). *Teaching core practices in teacher education*. Harvard Education Press.

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

Hiebert J. & Morris A.K. (2012). Teaching, rather than teachers, as a path toward improving classroom instruction. *Journal of Teacher Education*. *63*(2):92–102. https://doi.org/10.1177/0022487111428328

Hollins, E. (Ed.) (2015). *Rethinking field experiences in preservice teacher education: Meeting new challenges for accountability*. Routledge.

Janssen, F., Westbroek, H., & Doyle, W. (2014). The practical turn in teacher education: Designing a preparation sequence for core practice frames. *Journal of Teacher Education, 65*(3), 195–206. https://doi.org/10.1177%2F0022487113518584

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. T., & Wools, S. (2019). Perspectives on the validity of classroom assessments. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 8–23). Routledge. https://doi.org/10.4324/9780429507533

Kavanagh, S. S., Monte-Sano, C., Reisman, A., Fogo, B., McGrew, S., & Cipparone, P. (2019). Teaching content in practice: Investigating rehearsals of social studies discussions. *Teaching and Teacher Education, 86*, 102863. https://doi.org/10.1016/j.tate.2019.06.017

Kazemi, E., Ghousseini, H., Cunard, A., & Turrou, A. C. (2015). Getting inside rehearsals: Insights from teacher educators to support work on complex practice. *Journal of Teacher Education, 67*(1), 18–31. https://doi.org/10.1177/0022487115615

Liu, S., Bell, C. A., Jones, N., & McCaffrey, D.F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability, 31*(1), 31–61. https://doi.org/10.1007/s11092-018-09291-3

Martin-Raugh, M. P., Reese, C. M., Tannenbaum, R. J., Steinberg, J. H., & Xu, J. (2016a). Investigating the relevance and importance of high-leverage practices for beginning elementary school teachers (Research Memorandum No. RM-16-11). Educational Testing Service.

Martin-Raugh, M. P., Reese, C. M., Phelps, G. C., Tannenbaum, R. J., Steinberg, J. H., & Xu, J. (2016b). Investigating the relevance and importance of English language arts content knowledge areas for beginning elementary school teachers (Research Memorandum No. RM-16-08). Educational Testing Service.

Martin-Raugh, M. P., Reese, C. M., Howell, H., Tannenbaum, R. J., Steinberg, J. H., & Xu, J. (2016c). Investigating the relevance and importance of mathematical content knowledge areas for beginning elementary school teachers (Research Memorandum No. RM-16-10). Educational Testing Service.

Matsumoto-Royo, K., & Ramírez-Montoya, M. S. (2021). Core practices in practice-based teacher education: A systematic literature review of its teaching and assessment process. *Studies in Educational Evaluation, 70*, 101047. https://doi.org/https://doi.org/10.1016/j.stueduc.2021.101047

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, January). High-leverage practices in special education. Council for Exceptional Children & CEEDAR Center.

Mikeska, J. N., & Howell, H. (2020). Simulations as practice-based spaces to support elementary teachers in learning how to facilitate argumentation-focused science discussions. *Journal of Research in Science Teaching, 57*(9), 1356–1399. https://doi.org/10.1002/tea.21659

Mikeska, J. N., Howell, H., & Straub, C. (2019). Using performance tasks within simulated environments to assess teachers' ability to engage in coordinated, accumulated, and dynamic (CAD) competencies. *International Journal of Testing, 19*(2), 128–147.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–67.

Moss, P. (2011). Analyzing the teaching of professional practice. *Teachers College Record, 113*, 2878–2896.

Pressley, T., Roehrig, A.D., & Turner, J.E. (2018) Elementary teachers' perceptions of a reformed teacher-evaluation system, *The Teacher Educator, 53*(1), 21-43. https://doi.org/10.1080/08878730.2017.1391362

Qi, Y., & Sykes, G. (2016). Eliciting student thinking: Definition, research support, and measurement of the *ETS® National Observational Teaching Examination (NOTE) Assessment Series* (Research Memorandum No. RM-16-06). Educational Testing Service.

Resnick, A.F., & Kazemi, E. (2019). Decomposition of practice and an activity for research-practice partnerships. *AERA Open, 5*(3), 1–14. https://doi.org/10.1177/2332858419862273

Self, E. A., & Stengel, B. A. (2020). *Toward anti-oppressive teaching: Designing and using simulated encounters*. Harvard Education Press.

Shaughnessy, M., & Boerst, T. A. (2018). Uncovering the skills that preservice teachers bring to teacher education: The practice of eliciting a student's thinking. *Journal of Teacher Education, 69*(1), 40–55. https://doi.org/10.1177/0022487117702574

Shaughnessy, M., Boerst, T. A., & Farmer, S. O. (2019). Complementary assessments of prospective teachers' skill with eliciting student thinking. *Journal of Mathematics Teacher Education, 22*(6), 607–638. https://doi.org/10.1007/s10857-018-9402-x

Stickler, L., & Sykes, G. (2016). Modeling and explaining content: Definition, research support, and measurement of the *ETS® National Observational Teaching Examination (NOTE) Assessment Series* (Research Memorandum No. RM-16-07). Educational Testing Service.

Stroupe, D., Hammerness, K., & McDonald, S. (2020). *Preparing science teachers through practice-based teacher education*. Harvard Education Press.

TeachingWorks. (2020). *High leverage practices.* https://library.teachingworks.org/curriculum-resources/high-leverage-practices/

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New Teacher Project.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education, 96*, 878–903. https://doi.org/10.1002/sce.21027

Witherspoon, M., Sykes, G., & Bell, C. A. (2016). *Leading a classroom discussion: Definition, supporting evidence, and measurement of the ETS® National Observational Teaching Examination (NOTE) Assessment Series* (Research Memorandum RM-16-09). Educational Testing Service.

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

Appendix A

Leading Group Discussion

Task Directions

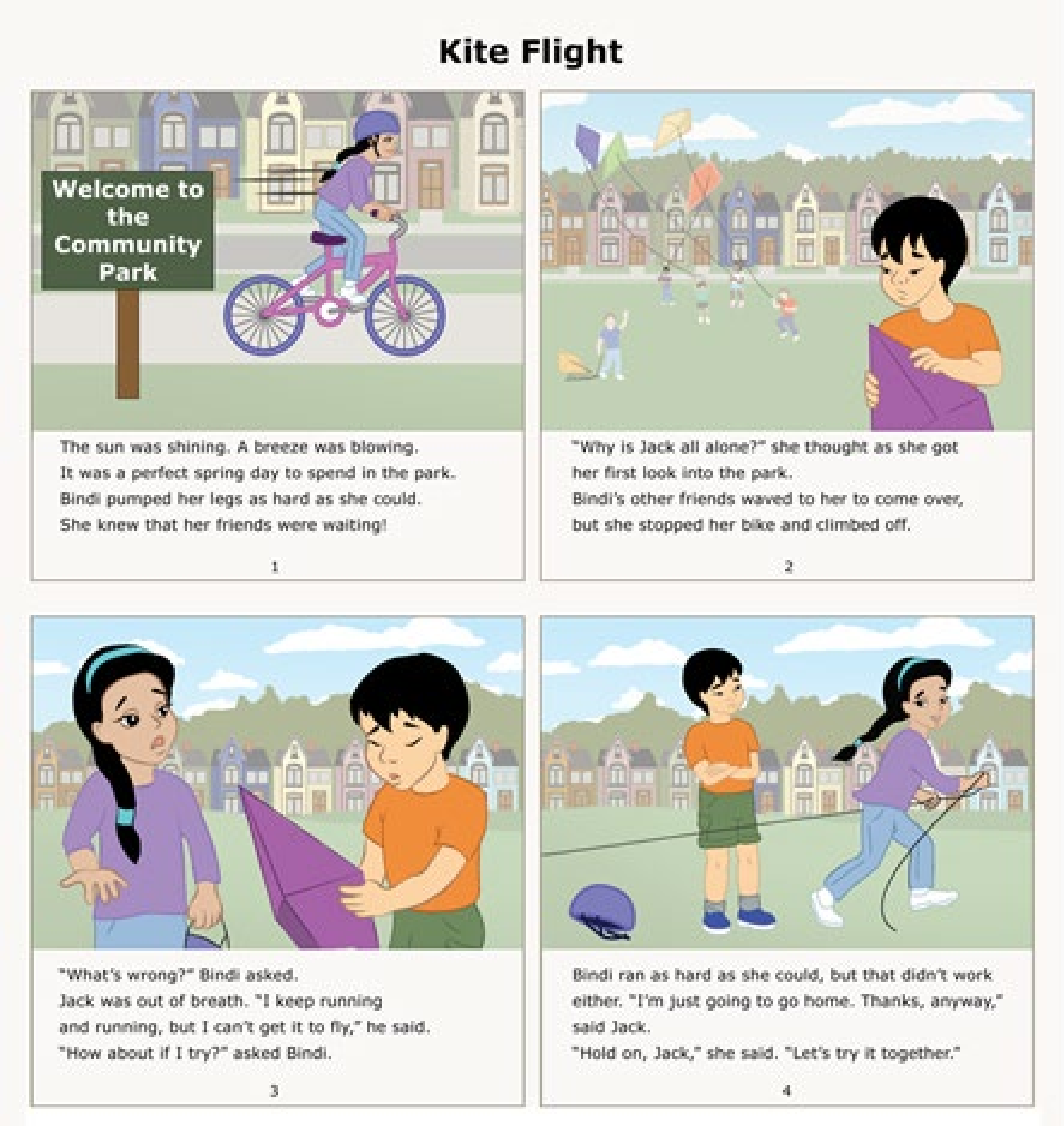
Grade level	First grade
Content area	Reading and language arts
Materials	A story called "Kite Flight," which will be displayed for you in two places. o *Discussion Materials—Prep* is a page to use during preparation. You can mark up or take notes on this page. You can refer to these notes during your performance, but the students will not be able to see this page. o *Discussion Materials* is a page to use during your performance. You and the students can see and write on this page during your performance.
Synopsis of the first part of the lesson	In the first part of the lesson, you led these first-grade students through a reading of a story called "Kite Flight." You ensured that all students understood the basic narrative of the text (i.e., what happened in the story), but you did not begin any further discussion about the text. To prepare students for the work they are about to do, you have already introduced the process of making inferences and explained that making inferences involves using evidence from the text and background knowledge to understand ideas not explicitly stated in the text.

Plan for this part of the lesson	In this next part of the lesson, you will lead a discussion aimed at helping students develop their ability to make and support inferences. The purpose of the discussion is to probe student thinking by having the students support their thinking with evidence from the text and by having them respond to each other's ideas. During this part of the lesson, students have copies of the text. You have asked the students to think on their own about Bindi's personality and what evidence they could use from the text (e.g., key events, characters' actions) to support different claims about her personality. Now you will begin a discussion about the students' ideas. The learning goal for the discussion is listed below. • In order to more deeply comprehend the text, make inferences about Bindi's personality and support those inferences with evidence from the text. Once your session begins, you should immediately launch into the discussion as if you had already carried out the first part of the lesson as described in the previous section. To signal to the students that you are ready, begin the discussion by saying, **"You all have done a good job reading the story 'Kite Flight.' Now I want us to think more carefully about Bindi's personality—about what she is like as a person. This discussion will give you the chance to think about your classmates' ideas and help you to develop and support your own ideas based on evidence from the story. So, what words would you use to describe Bindi? Who would like to start?"** During the discussion, you may use the *Classroom Materials* pages to record students' ideas and other information that may help to achieve the learning goals.

Additional Information

The following materials are designed to help you understand the reading and language arts content and the ways in which students at this grade level would be likely to interact with this text, including difficulties students might face with this content. The following can be used as a resource for you when planning the discussion, but this content is not designed for use with students and should not be used as a lesson plan.

Notes on the Text	The story "Kite Flight" is about a girl who is kind. The fact that she stops to help Jack instead of having fun with her friends shows that she is kind and generous because she puts Jack's feelings ahead of her own desire to have fun. Other personality traits that are supported by the text include, but are not limited to, the following. • Good friend • Nice • Smart • Creative • Inventive • Determined
Common Student Challenges	Students may struggle with the following. • Making inferences • Identifying appropriate evidence • Logically connecting and/or explaining evidence and inferences (e.g., explaining how coming up with multiple ideas is creative) • Distinguishing facts about Bindi from personality traits (e.g., "she seems to have lots of friends" versus "she is friendly and kind") • Vocabulary (e.g., students default to referring to Bindi as "nice")
Content Knowledge for Teaching	When readers make inferences they use clues in the text and their own background knowledge to come to a logical conclusion about something that is not explicitly stated. In this task, you will ask students to make inferences about personality traits. Personality traits are qualities of a person's nature that become evident in various ways that include, but are not limited to, what the person says and does.



Kite Flight

The sun was shining. A breeze was blowing.
It was a perfect spring day to spend in the park.
Bindi pumped her legs as hard as she could.
She knew that her friends were waiting!

1

"Why is Jack all alone?" she thought as she got
her first look into the park.
Bindi's other friends waved to her to come over,
but she stopped her bike and climbed off.

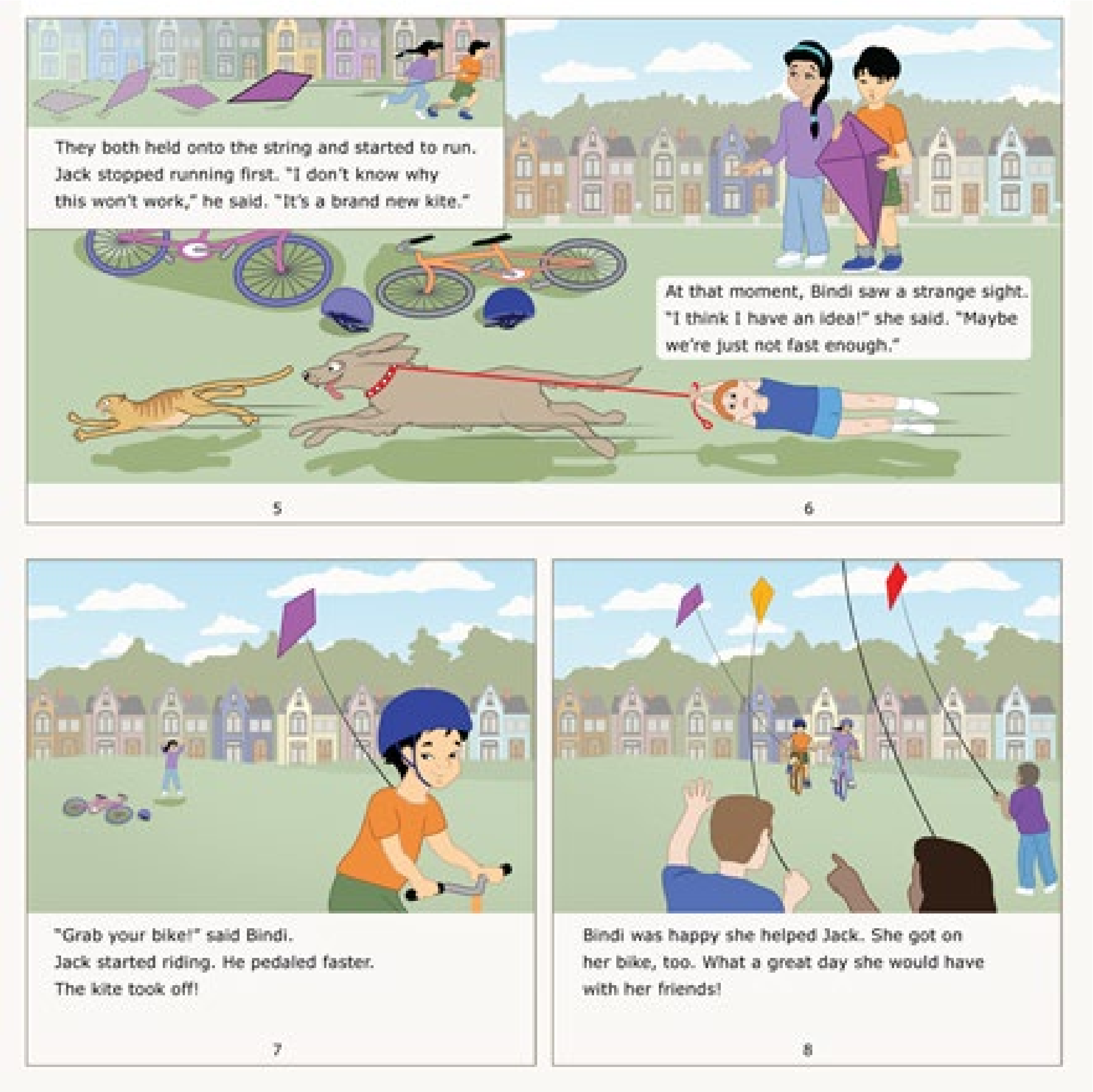
2

"What's wrong?" Bindi asked.
Jack was out of breath. "I keep running
and running, but I can't get it to fly," he said.
"How about if I try?" asked Bindi.

3

Bindi ran as hard as she could, but that didn't work
either. "I'm just going to go home. Thanks, anyway,"
said Jack.
"Hold on, Jack," she said. "Let's try it together."

4



They both held onto the string and started to run. Jack stopped running first. "I don't know why this won't work," he said. "It's a brand new kite."

At that moment, Bindi saw a strange sight. "I think I have an idea!" she said. "Maybe we're just not fast enough."

5

6

"Grab your bike!" said Bindi. Jack started riding. He pedaled faster. The kite took off!

7

Bindi was happy she helped Jack. She got on her bike, too. What a great day she would have with her friends!

8

Appendix B

***Table B1*. Descriptive Statistics by Task**

Task Type and Task	N	Min	Q25	Median	Q75	Max
Eliciting Math						
Eliciting_LM02	105	4	6	7	8	10
Eliciting_LM09	99	4	5	6	7.5	12
Eliciting_UM04	105	4	6	7	8	12
Eliciting_UM10	105	4	5.5	7	8	12
Eliciting ELA						
Eliciting_LE03	105	4	6.5	7.5	9	11
Eliciting_LE07	105	4	6.5	7.5	8.5	11
Eliciting_UE04	99	4	5.75	7	8.5	12
Eliciting_UE08	105	4.5	6.5	7.5	8.5	10.5
Discussion Math						
Discussion_LM02	105	4	5.5	6.5	7	10
Discussion_LM05	105	4	6	7	7.5	11
Discussion_UM01	105	4	6	7	8	11
Discussion_UM06	99	4	5.5	7	7.5	10.5
Discussion ELA						
Discussion_LE04	99	4	6.25	7	8	11
Discussion_LE06	105	4	6	7	8	11.5
Discussion_UE03	105	4	5.5	6.5	7.5	9.5
Discussion_UE05	105	4	6	6.5	7.5	10.5
Modeling Math						
Modeling_LM06	105	4	5	7	8.5	11
Modeling_LM08	99	4	4.75	6.5	8	12
Modeling_UM01	105	4	5	6.5	8	11.5
Modeling_UM07	105	4	5	7	8	11
Modeling ELA						
Modeling_LE09	105	4	6	7.5	9.5	12
Modeling_LE13	105	4	5.5	7	9	11.5
Modeling_UE02	105	4	5	6.5	8.5	12
Modeling_UE18	99	4	5.5	8	9.5	12

Note: After each task is a four-character indicator of the task. The first letter is upper or lower grades (U or L) for K–2 or 3–6. The second letter is for the subject matter—mathematics (M) or ELA (E). And the last two digits are for the task number.