VALIDATING DECOMPOSITION OF A TEACHING PRACTICE FOR FORMATIVE ASSESSMENT FEEDBACK

Daniel Heck
Horizon Research, Inc.
dheck@horizon-research.com

Evelyn M. Gordon
Horizon Research, Inc.
egordon@horizon-research.com

Meghan Shaughnessy
Boston University
mshaugh@bu.edu

Anna Bruce
Horizon Research, Inc.
a.bruce@horizon-research.com

Tim Boerst
University of Michigan
tboerst@umich.edu

Nicole Garcia
University of Michigan
nmgarcia@umich.edu

SimulaTE is studying teaching simulations as formative assessments of pre-service teachers' (PST) practice of eliciting and interpreting students' mathematical thinking. Preparation and protocols that promote reliability and validity of the simulations as formative assessments will enhance their effectiveness and generalizability. Teacher educators who use the simulations document each PST's performance to generate feedback for the PST in nine categories, arising from a decomposition of the teaching practice into specific component skills or actions. A series of coordinated validation studies include research to determine if the nine categories are distinguishable through the use of the simulation assessments, and can benefit from attention beyond other experiences PSTs have in their teacher preparation programs.

Keywords: Assessment, Mathematical Knowledge for Teaching, Preservice Teacher Education, Teacher Educators

Framing and Purpose of the Study

Ideally, teacher preparation develops candidate's skills and abilities for ambitious instruction that promotes student learning and counters inequities in outcomes. We ground our work in the understanding that frequent opportunities to engage in core practices of teaching, with formative feedback, can develop the knowledge, skills, and dispositions necessary for nurturing young learners of mathematics. Formative assessment provides pre-service teachers (PSTs) with feedback to improve their practice (Grossman, 2010), which is considered crucial for teacher preparation (Darling-Hammond et al., 2005; AMTE, 2017). It requires teacher educators to see teaching practices in action, yet traditional field settings afford neither frequent accessibility nor opportunities for deliberate work on specified facets of teaching. Simulations of mathematics teaching practices are an approximation that can provide early, frequent, and substantive formative assessment opportunities while engaging PSTs in particular facets of teaching.

PSTs begin preparation with views on teaching that need to be surfaced and, in some cases, challenged (Boerst et al., 2020; Shaughnessy & Boerst, 2018; Shaughnessy et al., 2020). Work initiated at the University of Michigan has produced multiple simulations to engage and refine PSTs' practice of eliciting and interpreting students' mathematical thinking. By revealing PSTs' knowledge, skills, and dispositions and providing immediate feedback, the simulations are designed to facilitate growth (Shute, 2008; Hattie & Timperley, 2007). This study's dual purposes are to investigate the decomposition of the teaching practice into measurable components for providing feedback, and to consider whether these skills or actions can benefit from concerted attention beyond other experiences typical to teacher preparation programs.

This study is one of a series of studies generating evidence regarding validity arguments (AERA et al., 2014; Kane, 2001; 2013) for using the simulations as formative assessments. It focuses on two sources of validity evidence (AERA et al.): internal structure (specifically test component interrelationships) and relations to other variables. Initial evidence for two specific claims of the validity argument are addressed here: (1) the nine component skills/actions of the teaching practice can be measured distinctly through simulation performances so that feedback can be specifically targeted, and (2) PSTs' other experiences in teacher preparation do not fully develop the component skills/actions of the teaching practice.

Teaching Simulations as Formative Assessments

Using the teaching simulations as formative assessments involves three interacting roles:

- The PST prepares for, engages in, and debriefs what they learn via the teaching practice of eliciting and interpreting student thinking with a Simulated Student.
- The Simulated Student is an adult prepared to follow a provided profile and to respond in specific ways to anticipated questions and prompts. (Student role)
- The Teacher Educator (TE) documents the PST's performance and provides formative feedback based on the performance. (Proctor role)

Figure 1 illustrates the full formative assessment process. The underlined components in the figure indicate the parts of the process investigated in this part of the validity studies.

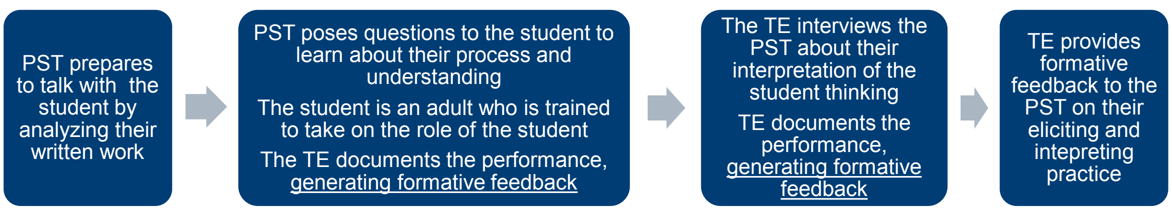
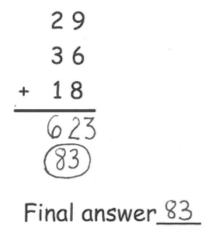


Figure 1: Structure of Teaching Simulations as Formative Assessments

The tasks in the simulations represent core content of elementary mathematics. The student work and specifications of the student role are evidence-based recreations of student thinking about that content (Shaughnessy et al., 2012). Figure 2 illustrates key elements of an assessment.

Mathematics topic: Multi-digit addition

- **The student's process:** The student is using the column addition method for solving multi-digit addition problems, the student is working from left to right.
- **The student's understanding of the ideas involved in the problem/process:** The student has conceptual understanding of the procedure including why combining is necessary (and when and how to combine).
- **Other information about the student's thinking, language, and orientation in this scenario:** The student talks about digits in columns in terms of the place value of the column. The student uses the term "combining" to refer to trading/carrying/regrouping.



Sample PST prompts	Sample Responses
What did you do first?"	"I added the tens: 2 + 3 + 1 and I got 6."
"How did you get from 623 to 83?"	"I had to combine the 6 and the 2."
"Why did you need to combine those numbers?"	"Because they're both tens."

Figure 2: Excerpts from a Sample Teaching Simulation Protocol

The content of the student work in the assessments was purposefully selected to cover mathematics concepts that PSTs are expected to have a strong understanding of and to provide insight into their capabilities. The simulation protocols were designed to reflect non-traditional approaches to solving mathematical problems or student thinking that results in an "incorrect" answer. The four simulation assessments used in this study included:

- Column Addition (CA): As shown in Figure 2
- Common Denominator: Comparing fractions, with an error in creating an equivalent fraction to compare using common denominators
- Common Numerator Correct: Comparing fractions, creating an equivalent fraction to compare using common numerators
- Expand and Trade: Multi-digit subtraction by writing quantities in expanded form and making trades among values before subtracting by place value, with an error in recording the value of a traded quantity

In teaching, "teachers pose questions or tasks that provoke or allow students to share their thinking about specific academic content in order to evaluate student understanding, guide instructional decisions, and surface ideas that will benefit other students" (TeachingWorks, 2024). The work of eliciting student thinking is conceived as: (a) formulating and posing questions to elicit and probe student thinking; (b) listening to and interpreting how students respond; (c) developing additional questions or tasks to pose; and (d) making sense of what students know and can do. Interpreting students' thinking is integral to eliciting, but is a distinct, overarching practice relying on broader information. It is conceived as: (a) sampling from evidence of student thinking and (b) using insight to articulate inferences grounded in the evidence. These practices take place within and across lessons, and in longer cycles of teaching that depends on learning about students to drive instruction (TeachingWorks, 2024).

Drawing on these conceptualizations, the simulation assessment situation and its documentation are based on a the following decomposition of the teaching practice into nine component skills or actions.

- Eliciting Process (EP): Uses questions or prompts to the student regarding their process for solving the task
- Interpreting Process (IP): Describes the student's process for solving the task
- Probing Understanding (PU): Uses questions or prompts to the student regarding their understanding of the mathematics of their process
- PST-Generated Interpretations of Student Understanding (PGSU): Spontaneous description of the student's understanding of the mathematics of their process
- Prompted Core Interpretations of Student Understanding (PCSU): Prompted description of core elements of the student's understanding of the mathematics of their process
- Attending to Student Thinking (ST): Asking questions about the written work and attending to what the student says in response to questions
- Applying Mathematics Knowledge for Teaching (MKT): Generating a task that can be used to confirm PST's understanding of the student's process
- Using Mathematics Knowledge and Skill (MKS): Applying the student's process to a new example, Generalizing about the mathematics/reasoning of the student's process
- Respecting the Student and Their Thinking (RS): Interacting with the student, and describing their work in ways that respect them as learners/knowers/doers of mathematics

To ensure assessment evidence about these specific components arises, the simulated student will disclose aspects of their process and understanding only when the PST deliberately prompts for it. Similarly, the PST is asked during the debriefing interview to recount very specifically what they learned about the student's process and understanding, supporting their claims with evidence they gathered. To further assess their application of mathematical knowledge for teaching and use of mathematics knowledge and skill regarding the targeted content, the PST is also asked to generate a problem to confirm what they learned about the student's process, and to explain the mathematics ideas undergirding the student's process and understanding. An online tool with protocols specific to each assessment (about 75 items) supports the teacher educator in documenting this fine-grained information. This documentation generates a level of performance (1-4) and formative feedback for each of the nine components. The TE can then use the performance levels and feedback to guide a discussion with the PST about areas of strength and potential improvement. The nine components are:

Study Methods and Participants

Data Collection

Data to address the two claims of the validation argument were collected between April 2023 and March 2024. Assessments were administered to 200 PSTs at 14 higher education institutions. Demographic data on participants are shown in Table 1.

Table 1: Characteristics of the Participants

Characteristic (N)	Percent of Respondents
Educational Attainment (199)	
Undergraduate student	96
Undergraduate degree in education or STEM discipline	4
Student Teaching or Internship (199)	
Had not yet begun student teaching or internship	80
Was currently doing student teaching or internship	20
Sex (199)	
Female	93
Male	5
Non-binary/non-conforming	1
Prefer not to answer	1
Hispanic or Latino (199)	
Yes	13
No	87
Race† (199)	
White	87
Asian or Asian American	4
Black or African American	3
American Indian, Native American, or Alaskan Native	1
Native Hawaiian or Pacific Islander	1
Prefer to self-describe	1
Prefer not to answer	1
Age (184)	
Traditional undergraduate-aged student (born 1998 to 2005)	96
Non-traditional undergraduate-aged student (born 1980 to 1997)	4

† Respondents were allowed to select more than one option; therefore, percent of respondents may add to more than 100.

Seven researchers, including authors 1, 2, and 4, prepared to administer the four assessments by learning the student and proctor roles and documenting performances in sample videos. The research team established reliability in both administration and documentation (Boerst et al., 2023; Heck et al., 2023). Researchers were assigned in multiple pairings to conduct site visits for data collection. A pair administered two assessments to each PST, alternating to distribute who served in the student and proctor roles. Each PST completed two of the four assessments, purposefully assigned to ensure equal distribution of assessments. Column Addition was administered to 90 PSTs, Common Denominator to 106, and Common Numerator Correct and Expand and Trade to 102 each.

Researchers' documentation generated a level of performance (1-4) for each component, along with potential feedback for discussion. Descriptive results for performance level scores on the nine component skills/actions are presented in Table 2. For these studies, feedback was not shared or discussed with participants to ensure it did not influence their performance on the second assessment they completed.

Table 2: Component Performance Level Scores

Skill/Action	Min	Max	Mean	SD
EP	1	4	2.95	1.03
IP	1	4	3.12	0.82
PU	1	4	1.85	0.99
PGSU	1	3	1.88	0.79
PCSU	1	4	2.58	1.12
ST	1	4	3.01	0.22
MKT	1	4	3.49	0.89
MKS	1	4	2.89	1.19
RS	1	4	3.79	0.55

Data Analysis, Results, and Findings

The first claim of the validity argument: the nine component skills/actions of the teaching practice can be measured distinctly, was examined in this study using the levels of performance that the documentation tool generates. A lack of correlation among the nine scores would offer evidence supporting this claim. Table 3 summarizes, for the 36 possible combinations of components, the correlations that were statistically significant. All were positive.

Table 3: Significant Correlations Between Components by Simulation Assessment

Comp.	EP	IP	PU	PGSU	PCSU	ST	MKT	MKS
IP	CA CD							
PU								
PGSU	CD CN	CD CN	CA CN ET					
PCSU	CN	CD CN	CA CN ET	CA CD CN ET				
ST	CA CN							
MKT		CD CN		CA	CA			
MKS	CA CN	CA CD CN ET	CD	CD CN	CA CD CN		CA CD	
RS			CD					

No significant correlations were found for 17 combinations of components on any of the assessments, and 5 other combinations produced a significant correlation on only one assessment.

Significant correlations (ranging from 0.20 to 0.55) were found for at least one combination involving each component. However, the most common significant correlations involved one of four components: IP (5 combinations of components, 12 instances across assessments), PGSU (6 combinations, 14 instances), PCSU (6 combinations, 14 instances), or MKS (6 combinations, 14 instances). In fact, only two other combinations—EP with ST in two instances and PU with RS

Kosko, K. W., Caniglia, J., Courtney, S., Zolfaghari, M., & Morris, G. A., (2024). *Proceedings of the forty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Kent State University.

in one instance—produced significant correlations.

Overall, these results provide mixed evidence regarding the claim that the nine components can be measured distinctly via the assessments. The extent to which these components are related has implications for targeting feedback to inform improvement on each component. Four components appear to be related to multiple others, suggesting that providing feedback on these components may be especially important for developing capabilities with the overall practice. Moreover, feedback on these four components might become especially useful by discussing their relevance to other components or the overall practice.

The second claim of the validity argument: PSTs' other experiences in teacher preparation do not fully develop the component skills/actions, was examined by predicting performance levels for each component using information PSTs reported about their progress in their programs. Specifically, the analysis considered their concurrent enrollment (N=96) or completion (N=98) of a mathematics for teaching (MfT) course, as well as their completion of other mathematics courses that are foundational (e.g., college algebra; N=89) or advanced (e.g., calculus; N=108). It also examined their concurrent or completed engagement in a student teaching placement (N=40).

Table 4 summarizes the results of a set of HLM analyses (scores nested within PSTs) predicting the performance level score for each component using data on PSTs' experiences in their preparation programs. Since PSTs were assigned to different pairs of assessments, dummy codes were also included to control for which assessment produced each performance level score. A lack of predictive association between PSTs' experiences and the performance levels on the simulation assessments offers initial supporting evidence for this claim.

Table 4: Positive and Negative Effects of PST Experiences on Components

Experiences	EP	IP	PU	PGSU	PCSU	ST	MKT	MKS	RS
Foundational Math								Neg.	Neg.
Advanced Math									
No MfT			Neg.		Neg.				
Enrolled in MfT				Neg.			Neg.		
Completed MfT									
Student Teaching							Neg.		

For three of the nine component skills/actions, PSTs' experiences in teacher preparation programs did not predict performance level scores. Variations in performance level scores for five of the other components were each predicted by only one type of experience. The remaining component was predicted by two experiences. PSTs who were neither concurrently enrolled in nor had completed a Mathematics for Teaching course with PU (respectively, -0.86 points, $p=.008$; -1.04 points, $p=.002$) and PCSU (respectively, -1.27 points, $p<.001$; -1.39 points, $p<.001$) suggests that participation in such courses contributes to development of these components of the practice the simulation assessments address. However, very few students who participated in this study (N=5) fell into this category. It is likely that such courses are an early requirement in most elementary education programs, so when participants were considered eligible for this study, they were already enrolled in an MfT course. Two other components showed predicted differences in performance levels between students concurrently enrolled and those who had completed MfT courses: PGSU (-0.18 points, p=0.048) and MKT (-0.24 points, p=0.013). To the extent that this smaller distinction represents differences in progress through teacher preparation programs, the lack of prediction of performance level scores on most components between these two conditions lends support to the validity claim. Further research involving PSTs who have not yet enrolled in MfT classes would be worthwhile.

PSTs' completion of a Foundational mathematics course predicted lower performance level scores on MKS (-0.45 points, p=.010) and RS (-0.13, p=0.036). Rather than calling the validity claim into question, this negative association may suggest that students whose mathematics coursework in college includes foundational content are likely to need more help in using mathematics knowledge and respecting student thinking in their teaching practice. Further research to pinpoint why some PSTs complete these courses and whether it signals something about their general mathematics knowledge would be informative.

PSTs' concurrent engagement in student teaching predicted a lower performance score on MKT (-0.27 points, p=.025). Again, this negative association does not challenge the validity claim. Rather, it might suggest a need to further support PSTs in making use of mathematics knowledge for teaching when they are student teaching. Additional longitudinal research would be informative to understand if entry into student teaching somehow affects PSTs' ability or propensity to apply MKT in the practice of eliciting and interpreting student thinking.

On the whole, these results provide initial evidence that experiences in teacher preparation are not likely to fully develop PSTs' abilities in the teaching practice of eliciting and interpreting student thinking. By extension, the simulation experience and associated feedback on the component skills/actions appears to offer a unique opportunity to support PSTs in more fully developing their capabilities with this practice.

Conclusions and Next Steps

The mathematics preparation of elementary teachers should develop their capabilities to enact teaching practices that support young learners' growth in mathematical knowledge, fluency, and disposition. Coursework and field placements that traditionally make up the bulk of PSTs' experiences in teacher preparation provide opportunities for PSTs to develop foundational knowledge of mathematics and pedagogy and to learn about and engage in these practices to an extent. They do not offer early, frequent, and structured experiences for PSTs to apply what they are learning in low-risk, high-feedback settings to support improvement in their capabilities. Simulations designed for engagement in teaching practices not only offer early, frequent, and structured experiences, but provide a measure of authenticity of PSTs' performance of the practices and opportunities for teacher educators to give immediate feedback to inform learning and improvement (Boerst et al., 2020; Darling-Hammond et al., 2005; Grossman, 2010).

Teaching simulations are resource intensive to develop and time intensive to use for formative assessment in teacher education. Strong validity must undergird their use to justify these investments. Kane's (2001; 2013) recommendations for developing and testing a validation argument require that the specific claims underlying the processes for administering assessments, generating results, and using results be stated and studied. Validation, in this view, is an ongoing process of amassing evidence to support or refute and, if necessary, refine these claims.

Prior work has demonstrated that preparation and support provided in the assessment materials result in consistent enactment of the simulations (Boerst et al., 2023) and reliable documentation of performances (Heck et al., 2023). In this study, the generated results were examined to test two additional validity claims, that (1) the component skills/actions of the teaching practice can be measured distinctly, and (2) PSTs' other experiences in teacher preparation do not fully develop these skills/actions. Analyses of data collected from PSTs in multiple teacher education programs on four different simulation assessments provided evidence supporting both claims, along with some discrepant evidence to be further studied.

Next steps in this validation work include further study of the first claim through an exploratory factor analysis of the items used to generate the component performance level scores. These data will also support analyses to test two additional claims addressing the response process as a source of validity evidence (AERA et al., 2014), within the full validation argument. First, the distribution of the four assessments across PSTs in various programs and their planned administration by multiple researchers serving in the student and proctor role will support a variance components analysis to examine the claim that the performance levels scores are mainly due to variations in the performance and not due to the effects of the specific assessment or the individuals playing the student and proctor roles. Second, data from this process were gathered from back-to-back performances on simulation assessments without sharing or discussing the generated feedback in between. Other data gathered within the larger project offer cases of the same pairs of assessments being administered with sharing and discussion of the generated feedback. These two situations will be contrasted to study the claim that engaging with the generated feedback promotes learning and improvement in performance.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) Grant No. 2101343. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/uploads/7/6/6/4/76643089/ standards_2014edition.pdf

Association of Mathematics Teacher Educators. (2017). *Standards for Preparing Teachers of Mathematics*. Available online at amte.net/standards.

Boerst, T., Craven., L., Heck, D., Shaughnessy, M., & Garcia, N. (2023). *Validating simulations to support broader use as formative assessments*. Poster presented at the National Council of Teachers of Mathematics Research Conference, October 24-25, Washington, DC.

Boerst, T. A., Shaughnessy, M., DeFino, R., Blunk, M., Farmer, S. O., Pfaff, E., & Pynes, D. (2020). Preparing teachers to formatively assess: Connecting the initial capabilities of preservice teachers with visions of teaching practice. In C. Martin, D. Polly, & R. Lambert (Eds.), *Handbook of Research on Formative Assessment in Pre-K through Elementary Classrooms* (pp. 89–116). IGI Global. https://doi.org/10.4018/978-1-7998-0323-2

Darling-Hammond, L., Pacheco, A., Michelli, N., LePage, P., Hamerness, K., & Youngs, P. (2005). Implementing curriculum renewal in teacher education: Managing organizational and policy change. In L. Darling-

Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 442–479). John Wiley & Sons. https://doi.org/10.5860/choice.43-1083

Grossman, P. (2010). *Learning to practice: The design of clinical experience in teacher preparation*. AACTE & NEA policy brief.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heck, D., Gordon, E. M., Shaughnessy, M., Boerst, T., & Garcia, N. (2023). Reliability and validity of documenting elementary pre-service teachers' performance in a teaching simulation. In T. Lamberg & D. Moss (Eds.), *Proceedings of the Forty-fifth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 1), 563–567. University of Nevada, Reno.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

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

Shaughnessy, M., & Boerst, T. (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., Sleep, L., & Ball, D. L. (2012, April). *Exploring how the subject matters in pedagogies of practice*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

Shaughnessy, M., DeFino, R., Pfaff, E., & Blunk, M. (2020). I think I made a mistake: How do prospective teachers elicit the thinking of a student who has made a mistake? *Journal of Mathematics Teacher Education*. https://doi.org/10.1007/s10857-020-09461-5

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

TeachingWorks (2024). *Mathematics high-leverage practices: Eliciting and interpreting*. https://library.teachingworks.org/curriculum-resources/materials/mathematics-eliciting-and-interpreting