RELIABILITY AND VALIDITY OF DOCUMENTING ELEMENTARY PRE-SERVICE TEACHERS' PERFORMANCE IN A TEACHING SIMULATION

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

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 relevant feedback for the PST. As part of a coordinated set of validity studies, six researchers were prepared on the documentation protocol. Consistency of documentation within the group and with the simulation developers' judgments provided evidence supporting reliability and validity of the documentation protocol.

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

Framing and Purpose of the Study

Mathematics teacher preparation ideally produces skillful and capable professionals whose classroom teaching will promote ambitious goals for student learning and counter chronic disparities in educational outcomes for students. Achieving this goal requires early and frequent engagement in practices of teaching with formative feedback to develop sophisticated knowledge, skills, and dispositions necessary for nurturing young learners of mathematics.

Formative assessment is a crucial component in teacher preparation (Darling-Hammond et al., 2005; AMTE, 2017) because it provides pre-service teachers (PSTs) with feedback they need to improve their practice (Grossman, 2010). It requires seeing teaching practices in action, but traditional field settings are limited in terms of frequent accessibility and the opportunity to work on specified facets of teaching. Simulations are an approximation of practice that can provide early, frequent, and substantive formative assessment opportunities while engaging PSTs in selected mathematics teaching practices.

PSTs enter preparation with knowledge, skills, and dispositions toward teaching that need to be surfaced, refined, or in some cases, counteracted (Boerst et al., 2020; Shaughnessy & Boerst, 2018a; Shaughnessy et al., 2020). To this end, efforts at the University of Michigan have resulted in a library of teaching simulations for elementary PSTs. The underlying premise is that PSTs' learning will be enhanced by performances of teaching practices that reveal the current state of their knowledge, skills, and dispositions and informing actions that facilitate growth (Shute, 2008; Hattie & Timperley, 2007). This study's purpose is to investigate reliability (consistency) and validity (accuracy) in the process for documenting performances to generate timely, interpretable, and actionable feedback.

Teaching Simulations as Formative Assessments

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

1. 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.
2. The Simulated Student is an adult prepared to follow a provided profile and to respond in specific ways to anticipated questions and prompts.
3. The Teacher Educator (TE) documents the PST's performance during the simulation and debriefing interview and provides formative feedback based on the performance.

Figure 1 illustrates the full formative assessment process. Two components are underlined in the figure to indicate the parts of the process investigated in this reliability and validity study.

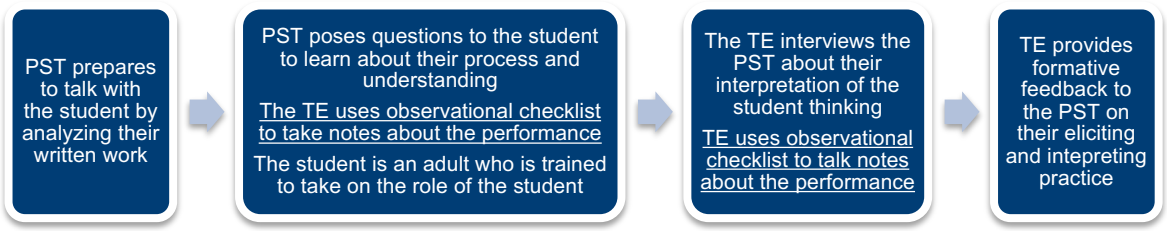
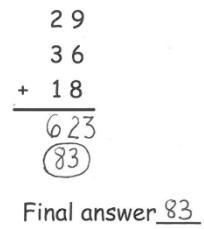


Figure 1: Structure of Teaching Simulations as Formative Assessments

Mathematics content in the simulations is high-leverage for elementary mathematics teaching (Shaughnessy et al., 2012) in that the tasks represent core disciplinary content and the student work and specifications of the role depict evidence-based recreations of student thinking about that content. Figure 2 is an excerpt of key elements of one simulation 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

By design, the student discloses aspects of their process and understanding only when asked. Consequently, the PST must prompt deliberately for each piece of information about the student's process and understanding. In the debriefing interview, the PST is asked to recount what they learned about the student's process and understanding in similar detail, including the particular evidence that supports their claims. In the interview, the PST is also asked about the

mathematical underpinnings of the student's process and understanding; their responses provide evidence of their mathematical knowledge for teaching (MKT) the targeted content.

Documentation of performances is undertaken at a very fine grain size to provide information about multiple components of the teaching practice. For each performance, documentation involves judgments for about 75 items, varying slightly depending on the particular content and task. Most judgments for the simulation documentation indicate whether or not the PST elicited or probed for specific pieces of information, and whether or not they took various actions (e.g., asking the student to write, posing a new problem) in doing so. Items regarding respect for the student's thinking are also included. In the debriefing interviews, items about process and understanding are judged in terms of whether correct/incorrect claims are offered and whether evidence for correct claims is provided or acknowledgement of lack of evidence and a proposed means of asking for it. The interview documentation includes judgments of correct or incorrect statements about the mathematics of the task and the student work. This documentation results in ratings for eight performance categories, listed in Table 1. These ratings generate feedback for the TE to discuss with the PST about strengths and areas for growth in their performance.

Reliability and Validity of Documenting PST Performances

Method for Producing Documentation Data

A crucial step in the process is documentation of the performance which links the enactment of the assessment to formative feedback. To use the simulation assessments effectively, TEs must consistently and accurately document performances so that the feedback is relevant to experience of the situation and tailored to support PSTs' growth. To examine this step, six mathematics education researchers (2 PhDs, 1 PhD student, 2 MAs, 1 BA) who work with multiple teacher education programs were prepared on the documentation procedure for four assessments (2 on multidigit operations, 2 on methods for comparing fractions). For each, the preparation included a meeting in which the developers introduced and provided practice with the content, student protocol, and documentation; independent documentation of performances from video samples of PST performances, then comparison and negotiation of judgments; and a follow-up meeting with the developers to discuss discrepancies and remaining questions.

The researchers next independently documented 16 videorecorded performances, four from each of the four assessments. These recordings were made as various members of the development group had enacted the simulation assessments in past years with 16 participants who were mostly PSTs and a few early career elementary teachers (to generate a range of performances). To mirror the expectation that the documentation occurs in real time, the researchers documented both the simulation and interview portions by watching the recordings uninterrupted. The researchers' documentations were then analyzed in two ways.

Method of Analysis

First, to investigate inter-rater reliability, researchers' documentations were compared within the group using Fleiss's Kappa, which measures the degree of agreement among multiple raters on multiple items, accounting for the probability of chance agreements. Values range from 0 (no agreement) to 1 (perfect agreement), with values above 0.6 considered substantial agreement, and above 0.8 near perfect agreement. See the third column of Table 1 for reliability results.

Second, to investigate validity, researchers' documentations were compared to a standard documentation, which was the collective judgment of the development team to document the same 16 performances. Unlike the research team, the development team had produced their documentation from multiple viewings and collective negotiation to determine appropriate documentation. In the validity analysis, each researchers' judgments were compared to the

developers' standard documentation using Cohen's Kappa. The six researchers' Cohen's Kappa values were then averaged. Ranging from -1 to 1, negative values represent greater non-agreement and positive values greater agreement. Values above 0.6 indicate substantial agreement and above 0.8 near perfect agreement. See the fourth column of Table 1 for results.

Results and Interpretation

Table 1: Fleiss's and Cohen's Kappa Values by Performance Category

Performance Category	Total Items	Fleiss's Kappa	Cohen's Kappa
Eliciting Process	28	0.86	0.86
Interpreting Process	38	0.86	0.86
Probing Understanding	26	0.75	0.72
Interpreting Understanding	112	0.83	0.78
Applying MKT	39	0.95	0.96
Other Math Knowledge/Skills	22	0.89	0.80
Attending to Student Thinking	16	0.90	0.94
Respecting the Student	16	0.96	0.90

As can be seen in the Fleiss's Kappa results in Table 1, inter-rater agreement among the six researchers was at least substantial (Landis & Koch, 1977) for all performance categories and can be considered near perfect for all but one. These results offer evidence that the researchers documented the performances quite similarly to one another, with the most common differences for Probing Understanding during the PST's interaction with the student.

Cohen's Kappa results in Table 1 indicate strong agreement (Landis & Koch, 1977) between researchers' documentation and the development team's standard judgments, characterized as at least substantial for all eight performance categories and near perfect for six. The most common differences between researchers and the development team were in Probing Understanding during the PST's interaction with the student and Intepreting Understanding in the interview.

Conclusions and Next Steps

The preparation researchers received and structured guidance the protocol provided appear sufficient to support documentation of performances that is reliable across different individuals for a variety of the simulation assessments. Additionally, the preparation and guidance supported valid judgments in documentation in that the researchers' and developers' judgments matched to a high degree. These results provide promise that teacher educators can take up these complex formative assessment tools for use in their own programs. The performance areas of lesser reliability and validity (Probing Understanding, Interpreting Understanding) suggest that preparation should emphasize and guidance highlight key decisions on items in these categories. Important next steps in validation studies of the simulation assessments will examine, via stimulated recall interviews, how TEs in the field document performances and investigation of the relevance of feedback generated from the assessments that TEs share and discuss with PSTs to support their growth in the teaching practice of eliciting and interpreting student thinking.

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

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

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

Landis, J. R. & Koch, G. G. (1977), The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

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