USING A FORMATIVE EVALUATION FRAMEWORK TO VALIDATE A TEACHING OBSERVATION TOOL

EL USO DE UN MARCO DE EVALUACIÓN FORMATIVA PARA LA VALIDACIÓN DE UNA HERRAMIENTA DE OBSERVACIÓN DE LA ENSEÑANZA

Kathleen Melhuish
Texas State University
melhuish@txstate.edu

Brittney Ellis
Texas State University
b_e107@txstate.edu

Alejandra Sorto
Texas State University
sorto@txstate.edu

Teaching observation protocols serve purposes beyond research, such as providing formative feedback for teachers' growth in their practice. Such observation tool usage requires different approaches for validation. For this reason, we developed the formative teaching evaluation framework adapted from the student-assessment literature. We used this framework as a guide toward collecting and organizing evidence of practitioners using the Math Habits Tool as a means for formative assessment of teaching. This report focuses on how we designed surveys and interviews to collect validity evidence related to the formative teaching evaluation intentions of the MHT. Overall, we established that the MHT was being used as intended by teachers and school leaders, and we provide details about the development and analysis procedures we took toward validating this observation tool for practitioner use.

Keywords: Measurement; Professional Development

Teaching observation tools are often positioned as both research tools and tools to provide formative feedback to support teacher reflection and growth (Boston et al., 2015). Yet, as noted by Yee et al. (2022), these protocols tend to focus on how "they are intended to be used to measure teaching behaviors" (p. 219) without attention to how they can be used for professional growth. Common observation protocols have focused validation efforts on measurement elements such as achieving inter rater reliability, establishing criterion-related validity, or validation through panels or experts (e.g., Gleason, et al., 2017; Hill et al., 2012; Schoenfeld, et al., 2018). While these traditional validation approaches are critical, we suggest there is also a need to explicitly collect validity evidence related to intended use beyond research contexts. That is, if these tools are developed to also provide formative feedback for teachers, then it is crucial that we attend to the viability of this usage type.

In this report, we present the *formative teaching evaluation* framework adapted from Nicol and Macfarlane-Dick's (2006) literature-based categories of formative assessment for students. We used this framework to provide a means to collect and organize evidence of an observation tool's formative evaluation usage. We share data from the Math Habits Tool observation protocol project, focusing on how surveys and interviews were used to collect validity evidence from users related to the formative teaching evaluation intentions of the instrument.

Formative Evaluation and Teacher Feedback

Schools have become increasingly data-driven (Marsh & Farrell, 2015) with teacher observation by school leaders or peers serving as the most common data source for evaluating teaching (Firestone & Donaldson, 2019; Steinberg & Donaldson, 2016). Such evaluation

provides summative information as well as a mechanism to encourage instructional growth; that is, "while observations were initially conceived as tools for evaluation, such protocols are now seen as key levers for the improvement of teaching" (Hill & Grossman, 2013, p. 372). Hill and Grossman have suggested that observation needs certain qualities to productively support instructional growth, such as content, observers, and time.

Instructional growth necessitates a reflection process that cannot be achieved through experience and training alone (Loughran, 2002). Productive reflection goes beyond summative lenses and requires nuanced, formative lenses for consideration of how instructional practice is working and for whom (Zeichener & Liston, 1996). Instructional practice needs to be decomposed from broad practice to specific practices (e.g., Grossman et al., 2009) and align with subject-specific goals. Evaluation focused on broad practice without specific ties to content has been found to have insufficient details needed to reflect on and change practice (e.g., Rigby et al., 2017). Hunter and Springer (2022) identified four critical observation feedback characteristics from the literature: "a) aligns with an improvement area, (b) discusses the feedback's evidential basis, (c) sets specific improvement goals, and (d) includes actionable next steps" (p. 380). That is, the quality of observation is about the content and whether the feedback provides a roadmap to be actionable toward change in practice. This aligns with Hill and Grossman's argument that a small grain-size is needed to support such actions and planning, a theme consistent with Firestone and Donaldson's (2019) larger literature synthesis.

One of the cross-cutting benefits of observation is creating a shared language for discussing ideas around instruction (Firestone & Donaldson, 2019); however, the relationship between the observer and teachers can have a strong impact on whether a particular observation is viewed as supportive for professional growth. In Paufler et al.'s (2020) study, they found that teachers diverged on whether they thought an evaluation system was primarily serving the purpose of judgment or formative feedback with attention to the way the system was communicated and how meetings with the observer played out. While the literature suggests that observation and feedback can provide teachers with support for instructional growth, the potential for formative evaluation is not always met by the incorporation of observation and feedback alone.

The Formative Teaching Evaluation Framework

Literature on formative assessment centers on assessing students (e.g., Black & Wiliam, 2010). To assess formative teaching evaluation goals, we developed a framework composed of formative assessment attributes. In prior work, we started creating this framework by adapting Nicol and Macfarlane-Dick's (2006) synthesis on formative assessment for students (Melhuish & Thanheiser). We suggest that formative assessment for teaching meets the following criteria:

1. helps clarify what quality instruction is (goals, criteria, expected standards).
2. facilitates the development of reflection on teaching.
3. delivers high-quality information to teachers about their teaching and student learning.
4. encourages dialogue around teaching and learning.
5. encourages positive motivational beliefs and self-esteem.
6. provides opportunities to close the gap between current and desired instructional practice.
7. provides information to observers (including school leaders) that can be used to help shape their support for teachers.

To meet these goals, formative assessment tools must reflect teachers' efforts at enacting critical components of teaching, provide details capturing the nature of these efforts, and offer

direction for change and growth. That is, observation needs to be specific, actionable, and provide a language that can support robust conversation and reflection.

Study Background and Methods

The Math Habits Tool

This study is situated in a larger project aimed at designing and validating the Math Habits Tool (MHT). The tool components overlay with the instructional triangle (Figure 1; Cohen et al., 2003; Hawkins, 2002). The components were developed iteratively by appealing to literature on student-centered classrooms focused on justifying and generalizing and professional development with in-service teachers and school leaders. It meets the recommendations of explicitly situating observed components in mathematical activity and using a small grain-size.

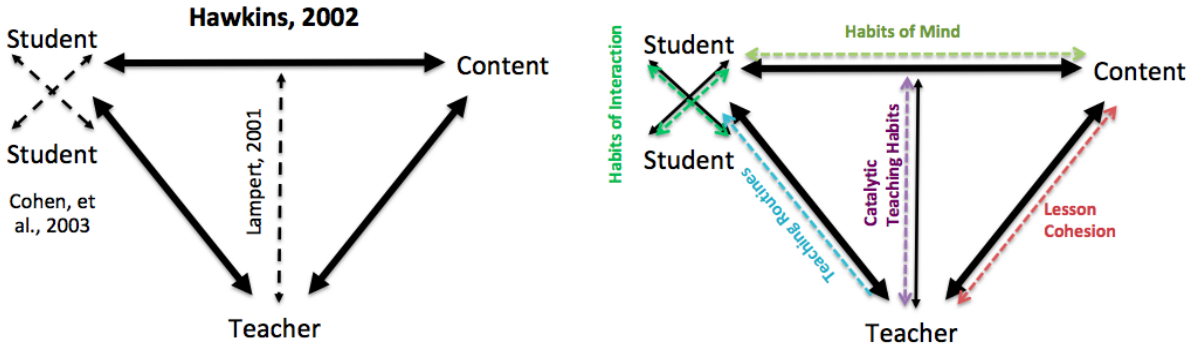

Figure 1. Instructional Triangle and Components of the MHT

Interpretation and Use Statement. Aligned with Carney et al. (2022), we provide the following interpretation and use statement. The Math Habits Tool measures standards-based instructional elements including teacher moves and routines and student math habits of mind and interaction that have been identified as important in the literature. Each component is measured via observation with a timestamp. The tool was designed for the full K-12 content spectrum for planning, observation, and reflection with teachers, principals, coaches, and within professional learning communities. It can be used to document an entire lesson or a brief drop-in segment. Rather than scores, the instrument provides timelines and frequencies for different activities. Analyzing patterns between students and teachers found in the timelines can be used to plan for teacher moves that may support increased student engagement in justification and generalization. Intended use is formative for planning, observing, and reflecting, not summative; that is, the certain occurrences or frequencies of activities should not be used to evaluate teachers.

Surveys, Interviews, Participants

The survey was designed to better understand how K-12 stakeholders (e.g., teachers, coaches, administrators/principals) the MHT, and verify whether the MHT is being used as a formative assessment tool (rather than for teacher evaluation). To address the second goal, we created multiple-choice survey items using the criteria for formative assessment for teaching framework as a guide. Two related versions were created and administered for teachers and school leaders, respectively. The survey included one open-response question intended to collect qualitative data about how practitioners perceived the purpose of the MHT. The survey was completed by 243 K-12 stakeholders (213 teachers, 30 school leaders) from a total of 53 different schools ranging from elementary to high school across the United States. All stakeholders had used the MHT and their experiences levels using the tool varied.

Survey Analysis. A principal component analysis was used to determine the underlying survey structure which led to the identification of three factors (elaborated in the results). We also considered correlation of all items to further bolster validity of the framework's construct coherence. The open-ended item responses regarding users' perceived purpose of the tool were qualitatively coded by research team members using an iterative process of open-coding, meeting to discuss codes, arriving at condensed themes, and applying these themes to the data.

Follow-Up Interviews. After survey analysis was completed, we interviewed 16 practitioners. A cluster analysis was conducted on the Likert scale responses to identify teacher profiles using a k-means clustering algorithm with a Euclidean distance metric. We ran the algorithm with two through eight clusters and used a D-B index to determine the best number of clusters. This analysis revealed five clusters (see Table 1). Participants who indicated agreement to be contacted for a follow-up interview were selected from each cluster, totaling 11 teachers. Five school leaders (principals/coaches) who volunteered were also interviewed. In the follow-up interviews, we asked practitioners to elaborate on their survey responses, share further information about their perceived purpose of the tool, and explain in detail how they used the tool with reference to particular parts (e.g., which parts were most used, which were least used) and contexts (e.g., who they used the tool with). A set of analytic notes were created after each interview to identify whether responses seemed consistent with surveys and how participants explained their use of the tool in terms of planning, teaching, and reflection.

Table 1: Description of the five clusters of survey respondents and dimensions of the survey

	Affective Items	Observation and Collaboration Items	Planning and Reflection Items
Cluster 1	Low	Low	Low
Cluster 2	Mid	High	Mid
Cluster 3	High	High	High
Cluster 4	Low	Low	Mid
Cluster 5	High	Low	Mid

Results

The survey results provided validity evidence that the MHT was meeting formative assessment goals for the majority of the respondents.

Surveys: Closed-Form Items

A scree plot indicated three distinct dimensions underlying the formative assessment survey (Figure 2). A set of sample items related to each dimension can be found in Table 2. Notice that individual activities were most prevalent with all teachers stating that they used the tool to plan for how to engage students in particular habits of mind and interactions. The observation and collaborative function occurred less frequently, yet indicated prevalent usage. This is not surprising or misaligned with intended usage because the time allotted for observation would be smaller than the potential tool usage as an individual reflection tool. Finally, we note that the MHT also seems to be supporting (at least to some degree) affective and understanding goals with majority of teachers agreeing that the tool supported instruction for justifying/generalizing and promoting confidence to refine instruction. We suggest that the closed-form results helped to bolster our validity of usage claims and provides a blueprint for other instrument developers.

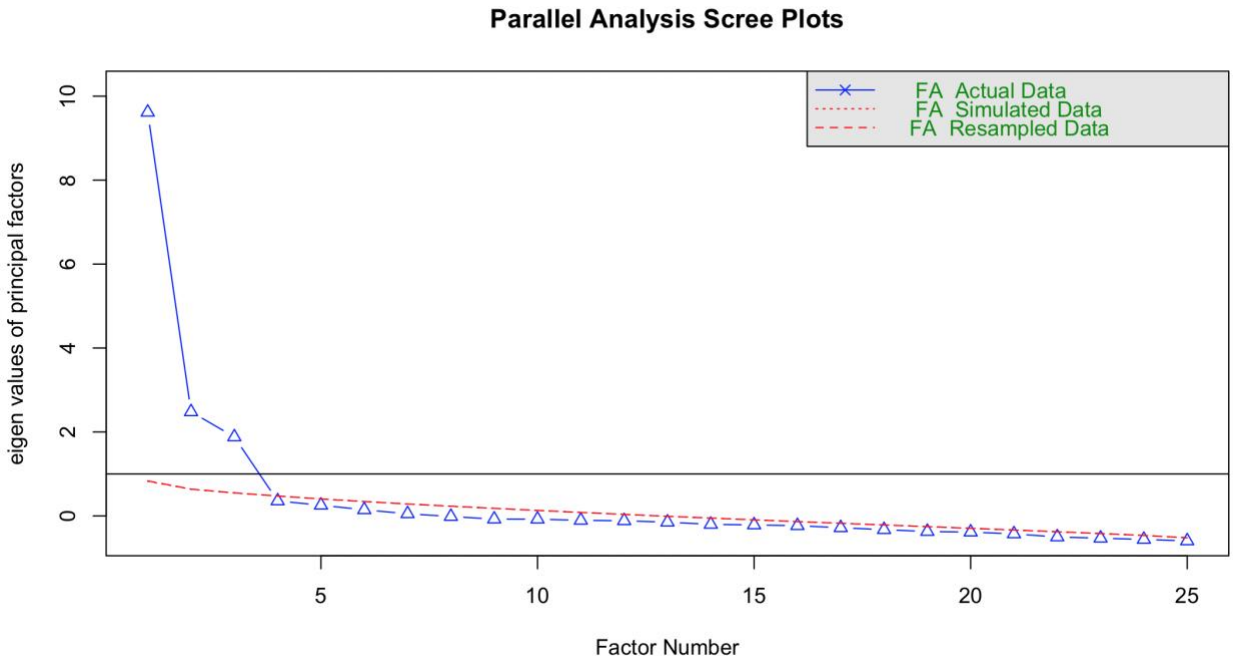

Figure 2: Scree plot. Note the three dimensions above the simulated data.

Table 2: Survey Items and Dimension

	Dimension	Percentages
I use the Math Habits Tool to plan how to engage students in habits of mind/interaction.	Individual	Never (0%) Only with PD (13%) Rarely/Occasionally (32%) Weekly/Daily (55%)
I use the Math Habits Tool to reflect on whether enacting teaching routines I planned to implement went as expected.	Individual	Never (4%) Only with PD (18%) Rarely/Occasionally (48%) Weekly/Daily (30%)
I use the Math Habits Tool to talk with school leaders (e.g., coaches, principals) about classroom observations.	Observation and Collaboration	Never (14%) Only with PD (25%) Rarely/Occasionally (59%) Weekly/Daily (2%)
I use the Math Habits Tool to observe teachers' classrooms with other teachers	Observation and Collaboration	Never (34%) Only with PD (34%) Rarely/Occasionally (31%) Weekly/Daily (1%)
By using the Math Habits Tool, I feel confident in refining my instruction.	Affect and Understanding	Strongly Disagree (3%) Disagree (9%) Agree (57%) Strongly Agree (31%)
Using the Math Habits Tool has helped me better understand how instruction can support students in justifying/generalizing.	Affect and Understanding	Strongly Disagree (3%) Disagree (3%) Agree (35%) Strongly Agree (59%)

Surveys: Open-Ended Item

The survey also asked teachers and school leaders to respond to the question: *In your own understanding, what is the purpose of the Math Habits Tool?* This question was used to better situate the users' responses on the closed-form items and gather additional evidence of how they conceptualized the tool's purpose. From the coding process, five main themes were identified: *common language and framework*, *planning and reflection on practice, observation and feedback, mechanism to promote students' mathematical reasoning,* and *a mechanism to support students' mathematical communication with each other.* Regarding common language and framework, stakeholders articulated that the MHT provided a specific language for both students and teachers to use together as well as a structure that reflects important elements of instruction. One teacher noted that the tool is a "quick reference for best learning/teaching practices in math," which can support student learning. A school leader further explained that the tool provides a "common language and conceptual grasp for a department of math teachers." This came up in 60% of the school leader responses (18 total) and 36% of the teacher responses (76 total).

The tool as a mechanism to promote students' mathematical reasoning was a salient theme for teachers, with 140 teachers (66%) noting this purpose in their responses. For example, a teacher wrote, "the purpose of the [tool] is to engage all students in mathematical ideas and understanding by using representations and making connections to build justifications and generalizations." While less of a focus for school leaders (10 total, 35%), many mentioned this alongside other purposes. For instance, one school leader stated the purpose of the tool was to "ground the work you do planning for lessons, discussions, tasks [and] connect things in a way that leads students to math conjectures and generalizations," which speaks to the tool as a framework, planning tool, and a tool to support discussions. A less salient theme for teachers was observation and feedback use, with 12 teachers (5.63%) and 13 school leaders (43.33%) mentioning this purpose, which aligns with our findings in the prior section.

Teachers disclosed either planning or reflecting on practice as the purpose of the tool; ideally, the goal would be for planning and reflection together to be viewed as a purpose. For instance, a teacher mentioned, "I use this to reflect on what was in my lesson, and I use this also to think about planning my upcoming lessons and activities." Overall, about 20% of teachers (41 total) and 27% of school leaders (8 total) mentioned planning and/or reflection on teaching. Finally, the tool as a mechanism to support students' mathematical communication with each other was brought up in 68 teachers' (32%) and 12 school leaders' (40%) responses. This was often talked about as facilitating discourse about mathematics in the classroom, to "engage students in purposeful conversations about math." Due to the nature of the question, practitioners may have thought to record only one purpose of the tool. For this reason, having practitioners expand on their responses or add other purposes in the interviews was critical for validation.

Interviews

Semi-structured interviews provided a space for open conversations with practitioners to collect additional evidence of how the tool was being used. It was a crucial part of our process to determine whether the tool was being used for formative assessment, and to confirm whether it was *not* being used for other purposes, such as summative evaluation. All interviewees expressed formative assessment uses with school leaders, as exemplified in the following quote from a K-5 mathematics coach:

> …sometimes the tool is more for me and then I just filter it [into] just general teacher language when I'm talking to the teacher, or sometimes I will say 'Okay let's look at the tool

> together,' since this is a teacher who is very familiar with it, and we can use that to do some of the reflection, but it's all really…how can I help this teacher to identify a goal for themselves to increase their practice.

This coach explained multiple purposes including common language, collaborating, reflecting, and planning. They went on to explain, "I think from the teacher's perspective… just the amount of formative assessment data you get from utilizing the tool with your students and just understanding how they're thinking, how they're making connections…." They voiced the collection and use of formative assessment that focused teachers' attention to what students are thinking. Another coach directly explained that they are "careful that… when we go in with admin it doesn't feel like another evaluation," recognizing that the tool should support learning goals but not summative evaluation.

In interviews, teachers also reflected that the MHT was not being used for summative evaluation. One 6th grade teacher elaborated that it is, "less of like a 'are you good, are you not?' Why is it good and just very vague. It makes it very concrete, easily talked about versus evaluative." Overall, teachers elaborated on the role the tool played in their teaching, their planning and reflection, and their collaboration in the school, and emphasized the communication role of the tool. For example, a high school teacher stated, "I also think it's a great onboarding tool for administrators who don't have experience in math classrooms specifically or even if they do, it's a great tool that launches conversation between teachers and administrators." They continued to discuss planning, saying, "But, like everything else about the planning with… those things in the tool in mind gave me life and joy and gave my students success and confidence and so that's why it's ingrained in me."

We purposefully selected teachers from clusters that had survey responses reflecting less achievement of formative assessment goals. From these interviews, we identified several ways teachers voiced that the MHT was not meeting formative assessment needs. One concern from a middle school teacher was a misalignment between the mathematics emphasized in the observation tool (such as justifying and generalizing) and the mathematics emphasized in the school's curriculum. She noted that the "philosophy of the curriculum is very rote math" and the tool was hard to use in the later parts of the year that were more focused on procedures. Another issue that arose was burn out, especially since the pandemic. This same teacher noted, "it was just like one more thing to make me feel like I was not a good teacher, and so I think that can be very, very discouraging." This comment provides negative evidence in relation to the affect and understanding dimensions (*encourages positive motivational beliefs and self-esteem* from the framework). In this case, she viewed the MHT as just one more thing to balance and imposing an ambitious image of instruction that was hard to enact.

Finally, some teachers noted that while they used the content of the tool in their thinking when planning, they no longer relied on the physical tool. A teacher in a 6th grade math and science Spanish immersion classroom stated that, "In my mind, it's a bridge to take us from traditional math classrooms where the teacher is delivering information, students are note taking, students regurgitate said information and we all move on and then students learn how to hate math." They continued explaining that "when I'm planning, I'm thinking about these things, but not through the [physical] tool." One explanation for this was that the school provided "lots of initiatives" and so they had other tools (besides the MHT) to provide the support they needed.

Discussion

As researcher tools become adapted to new contexts, it is important to validate for purposes beyond research. Our goal in this report was to share some of our efforts to validate an observation tool for a practitioner setting. To meet this aim, we developed the formative evaluation of teaching framework and a corresponding survey to administer to teachers and school leaders. Throughout this work, we take the general stance that tool validation should be done through the accumulation of different types of evidence to build a robust validity argument (Kane, 1992). While we have done substantial validation from a research perspective, in alignment with best practices suggested by other researchers, we found a need to consider the validity of the instrument for use by teachers and school leaders. That is, we wanted to lay out a systematic way to collect evidence of whether the MHT was successful in its formative evaluation intentions for practitioners. To do this, we collected evidence from many end users (teachers/school leaders) via a survey, and then additional evidence from detailed interviews with a targeted subset of practitioners who completed the surveys.

By using the formative evaluation of teaching framework as a guide while collecting evidence of the MHT's formative evaluation usage, we found that the MHT was being used as intended overall. The open-ended questions and interviews provided additional details and reflected some of the literature-based arguments for quality observation tools including providing a shared language (e.g., Firestone & Donaldson, 2019), tying into specific content (e.g., Rigby, et al., 2017), and providing actionable means to plan for and reflect on instruction (e.g., Hunter & Springer, 2022). We also explored how the same observation tool may not meet its formative assessment goals for all instructors. Our results indicated that there may be misalignment between curriculum and observation emphasis, competing school initiatives, and inferred quality evaluations that could interfere with the MHT's formative evaluation potential.

The ways in which school leaders use an observation tool as formative feedback may impact how teachers take up using the tool on their own and with other teachers. In the surveys and interviews, we noticed that teachers mentioned higher use of the MHT with professional development support. All the teachers and school leaders in our study had some professional development support around the use of the tool. Additional research is needed to see how the MHT formative evaluation use can be met with different degrees of support. Overall, we see this study as contributing an image of a positive step for validating tools, but additional work is needed for expansion to other contexts.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1814114. 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 National Science Foundation.

References

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. Phi delta kappan, 92(1), 81-90.

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. Mathematics Teacher Educator, 3(2), 154-175.

Cohen, D. Raudenbush, S. Ball, D. (2003). Resources, instruction, and research. Educational Evaluation and Policy Analysis, 25, 119-142.

Firestone, W. A., & Donaldson, M. L. (2019). Teacher evaluation as data use: what recent research suggests. Educational Assessment, Evaluation and Accountability, 31, 289-314.

Gleason, J., Livers, S., & Zelkowski, J. (2017). Mathematics classroom observation protocol for practices (MCOP2): A validation study. Investigations in Mathematics Learning, 9(3), 111-129.

Hawkins, D. (2002). The informed vision. New York, NY: Algora Publishing

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., ... & Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. Educational Assessment, 17(2-3), 88-106.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. Harvard educational review, 83(2), 371-384.

Hunter, S. B., & Springer, M. G. (2022). Critical feedback characteristics, teacher human capital, and early-career teacher performance: A mixed-methods analysis. Educational Evaluation and Policy Analysis, 44(3), 380-403.

Loughran, J. J. (2002). Effective reflective practice: In search of meaning in learning about teaching. Journal of teacher education, 53(1), 33-43.

Marsh, J. A., & Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making: A framework for understanding capacity building. Educational Management Administration & Leadership, 43(2), 269-289.

Melhuish, K. M., & Thanhesier, E. (2017). Using formative evaluation to support teachers in increasing student reasoning. In L. West & M. Boston (Eds.), Annual perspectives in mathematics education 2017: Reflective and collaborative processes to improve mathematics teaching (pp. 183–199). National Council of Teachers of Mathematics.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in higher education, 31(2), 199-218.

Paufler, N. A., King, K. M., & Zhu, P. (2020). Promoting professional growth in new teacher evaluation systems: Practitioners' lived experiences in changing policy contexts. Studies in Educational Evaluation, 65, 100873.

Schoenfeld, A. H., Floden, R., El Chidiac, F., Gillingham, D., Fink, H., Hu, S., ... & Zarkh, A. (2018). On classroom observations. Journal for STEM Education Research, 1, 34-59.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. Education Finance and Policy, 11(3), 340-359.

Rigby, J. G., Larbi-Cherif, A., Rosenquist, B. A., Sharpe, C. J., Cobb, P., & Smith, T. (2017). Administrator observation and feedback: Does it lead toward improvement in inquiry-oriented math instruction?. Educational Administration Quarterly, 53(3), 475-516.

Yee, S., Deshler, J., Rogers, K. C., Petrulis, R., Potvin, C. D., & Sweeney, J. (2022). Bridging the gap between observation protocols and formative feedback. Journal of mathematics teacher education, 25(2), 217-245.