Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1298

Facilitating Evidence-Based Instructional Coaching With Automated Feedback on Teacher Discourse

James Malamut
Stanford University

Dorottya Demszky
Stanford University

Christine Bywater
Stanford University

Michele Reinhart
Stanford University

Heather C. Hill
Harvard University

Instructional coaching often aims to ground teacher professional learning in classroom evidence, yet it is labor-intensive for coaches to obtain and curate such evidence. This study explores possibilities for utilizing an automated feedback tool to support evidence-based reasoning in coaching. We examined how mathematics coaches employed automated feedback within coaching, asking what affordances/limitations they anticipated, how they used the tool, and how they perceived it after initial use. We conducted semi-structured interviews with 20 coaches and recorded 41 coaching conversations made by nine of those coaches. Our analyses suggest that coaches used generated evidence to demonstrate effective practices, probe, challenge, and compare/generalize across moments. They valued the tool's efficiency, visual representations, and perceived objectivity but raised concerns about accuracy, particularly for student talk. The study highlights the importance of coach facilitation and suggests future research and design directions for the effective integration of automated tools in instructional coaching.

VERSION: September 2025

Suggested citation: Malamut, James, Dorottya Demszky, Christine Bywater, Michele Reinhart, and Heather C. Hill. (2025). Facilitating Evidence-Based Instructional Coaching With Automated Feedback on Teacher Discourse. (EdWorkingPaper: 25-1298). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xx9z-8f27

Facilitating Evidence-Based Instructional Coaching With Automated Feedback on Teacher Discourse

James Malamut [a], Dorottya Demszky [a], Christine Bywater [a], Michele Reinhart [a], Heather C. Hill [b]

[a] Stanford Graduate School of Education, 482 Galvez Mall, Stanford, CA 94305, USA

[b] Harvard Graduate School of Education, 13 Appian Way, Cambridge, MA 02138, USA

Author Note

CRediT Roles: James Malamut: Conceptualization, Investigation, Writing - Original Draft **Dorottya Demszky:** Conceptualization, Funding Acquisition, Writing - Review and Editing **Christine Bywater:** Data Curation, Project Administration **Michele Reinhart:** Data Curation **Heather Hill:** Conceptualization, Writing - Review and Editing

Acknowledgements: The authors are grateful to Janet Carlson, Arielle Boguslav, Jing Liu, Hannah Rosenstein, and all of the coaches and teachers who participated.

Funding: This work was supported by the National Science Foundation (#2337772).

Corresponding Author: James Malamut, Stanford University, Stanford, CA, USA

Email addresses: jmalamut@stanford.edu James Malamut, ddemszky@stanford.edu Dorottya Demszky, cbywater@stanford.edu Christine Bywater, michelereinhart@stanford.edu Michele Reinhart, heather_hill@gse.harvard.edu Heather Hill

Declaration of interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abstract

Instructional coaching often aims to ground teacher professional learning in classroom evidence, yet it is labor-intensive for coaches to obtain and curate such evidence. This study explores possibilities for utilizing an automated feedback tool to support evidence-based reasoning in coaching. We examined how mathematics coaches employed automated feedback within coaching, asking what affordances/limitations they anticipated, how they used the tool, and how they perceived it after initial use. We conducted semi-structured interviews with 20 coaches and recorded 41 coaching conversations made by nine of those coaches. Our analyses suggest that coaches used generated evidence to demonstrate effective practices, probe, challenge, and compare/generalize across moments. They valued the tool's efficiency, visual representations, and perceived objectivity but raised concerns about accuracy, particularly for student talk. The study highlights the importance of coach facilitation and suggests future research and design directions for the effective integration of automated tools in instructional coaching.

Facilitating Evidence-Based Instructional Coaching With Automated Feedback on Teacher Discourse

Robust empirical evidence suggests that instructional coaching can lead to improvements in both instruction and student learning (Kraft et al., 2018). Many successful programs center coaching conversations on classroom evidence, often by analyzing video of instruction, transcripts of teacher–student interactions, or student work (e.g., Allen et al., 2015; van Vondel et al., 2018). Yet despite documented benefits, evidence‑based reasoning is often underutilized in coach–teacher discussions due to competing demands for planning logistics, such as deciding when to teach different topics (Horn et al., 2017).

Video is a common coaching artifact that can productively anchor evidence-based reasoning (Amador et al., 2025; Gaudin & Chaliès, 2015). However, preparing video for coaching—recording, reviewing, clipping, and aligning it to goals—is time-intensive and resource-demanding, which limits routine use. Automated feedback tools may provide a complementary, more efficient, and multifaceted way of producing instructional artifacts. These tools can support the curation of relevant evidence to use in coaching conversations. Teacher–student talk ratios, distribution of participation, question types, and wait time can be highlighted, making abstract discourse features visible and comparable across different moments of lessons.

Despite this promise, we know little about how automated feedback actually functions within coaching practice. To examine its utility in evidence-driven coaching, we interviewed 20 mathematics instructional coaches and observed nine of them as they conducted coaching conversations with M-Powering Teachers, an automated feedback tool. We asked:

1. What affordances and limitations of an automated feedback tool do coaches anticipate?
2. How do they utilize automated feedback in coaching conversations?
3. How do they perceive an automated feedback tool after its initial use?

Literature Review

Evidence and Instructional Coaching

Instructional coaching can be an effective vehicle for teacher professional development, often leading to improvements in teaching, student achievement, or both (Campbell & Malkus, 2011; Desimone & Pak, 2017; Kraft et al., 2018; Russell et al., 2020). Some coaching activities, such as eliciting reasoning about observed instruction, may be especially productive (Kochmanski & Cobb, 2023). Reasoning using evidence from observed instruction allows teachers to "rethink their teaching, rather than merely extend their existing practices" (Horn et al., 2017, p. 51). Many instructional coaching programs use reasoning about evidence as a key feature of the experience. For example, the Mathematics Quality of Instruction coaching cycle requires teachers and coaches to "describe, without judgment, what happened" (Center for Education Policy, 2023, paragraph 2) in curated video clips. The Center to Support Excellence in Teaching (2021) includes analyzing the use of teaching practices with students as a core component of instructional improvement. However, the use of reasoning and evidence may be underutilized in teacher discussions about instruction (Horn et al., 2017; Lefstein et al., 2020).

Instructional artifacts such as observation notes, lesson plans, and video recordings are often used as sources of evidence during teacher professional learning. Classroom recordings capture granular classroom interactions—for example, teachers' questions and students' responses—which can then foster productive discussions among teachers (Borko et al., 2008) and shift teacher thinking and practice (van Es & Sherin, 2010). Amador et al. (2025) found that

incorporating video into coaching sessions increased how often mathematics teachers reasoned about the connections between specific elements of their instruction and evidence of impacts on student thinking. However, these studies also caution that the use of video (or any artifact) does not automatically lead teachers to use evidence or engage in reasoning. Instead, learning activities must be purposefully designed and facilitated to support teacher growth, for example by featuring activities that prompt reasoning and link to teachers' learning goals (Brophy, 2003; Erickson, 2014).

Although the selection of video clips can be a useful practice for coaches and teachers (Calandra et al., 2008), it can be very time-consuming and expensive. Further, individual video clips provide lots of rich information about particular moments in classrooms but do not necessarily support teachers in reasoning about connections between those moments (Coles, 2014). However, new technologies have emerged to help support the curation and analysis of instructional artifacts.

Automated Feedback as Instructional Artifacts in Coaching

To support coaches and teachers in analyzing classroom evidence, researchers have developed a variety of visualization and video annotation tools, including VideoNoter (Roschelle & Goldman, 1991), CORDTRA (Hmelo-Silver et al., 2008), Studiocode (Prusak et al., 2010), and KBDEX (Oshima et al., 2012). Such tools are designed to help teachers and coaches navigate and organize lengthy recordings by allowing playback and helping with the identification and tagging of salient patterns in the video. One of the most recent such tools, the Classroom Discourse Analyzer (CDA), integrates visualizations of classroom talk data (e.g., timelines comparing teacher vs. student talk; frequencies of certain "talk moves") with annotation and playback features (Chen et al., 2015). In a year-long randomized study, Chen et

al. (2020) found that the use of CDA prompted deeper teacher reflection and increased the use of productive teacher talk moves and improved student outcomes in contrast to a comparison group. Such tools can increase the productivity of video analysis, but they still typically rely on coaches viewing all or most of a lesson and then manually identifying noteworthy events such as specific teacher talk moves.

An emerging line of work explores the automation of classroom discourse analysis for the purposes of teacher professional learning. In fact, technological advances in automated speech recognition and natural language processing have spurred a new generation of tools that can automatically detect relevant moments based on classroom discourse. These tools, such as TeachFX (www.teachfx.com), TalkMoves (Suresh et al., 2021), and M-Powering Teachers (Demszky et al., 2023, 2025; Demszky & Liu, 2023), analyze recordings of teachers' instruction and provide evidence on topics such as teacher–student talk ratios and teachers' use of high-leverage talk moves—for instance, their uptake of student ideas or use of probing questions. Initially, these feedback tools were envisioned for teachers' self-directed professional learning (Demszky et al., 2025): Teachers uploaded lesson audio, privately reviewed and reflected on the generated feedback, set personal goals, and tracked their progress. As such, these tools were meant to provide personalized, evidence-based feedback to teachers who were otherwise reluctant to be observed (Little, 1990), thus serving as a low-cost supplement to group-based professional learning and an alternative to one-on-one coaching.

Early evidence indicates the potential of private, automated teacher feedback in improving instruction (Demszky et al., 2023, 2025; Jacobs et al., 2022, 2024). A randomized controlled trial with 224 K–12 math and science teachers showed that automated feedback increased teachers' use of focusing questions—prompts that probe student thinking—by 20%

after just two feedback cycles (Demszky et al., 2025). Likewise, a case study with 21 mathematics teachers by Jacobs et al. (2022) found that teachers valued automated feedback on dialogic talk moves and increased their use of such moves.

At the same time, these studies report key barriers to teachers' use of automated feedback tools. In interviews, teachers reported insufficient time for reflection, technical difficulties with feedback platforms, transcription errors that obscured classroom interactions, and struggles with interpreting quantitative data without support, each of which could undermine usability, trust, and effectiveness of the tools (Demszky et al., 2025; Jacobs et al., 2022, 2024). Likewise, Jacobs (2022) notes that teachers working independently without expert support often struggled "to distill what the feedback on specific talk moves actually meant for them" (p. 11).

Consequently, a valuable research direction emerging from this line of work is the integration of these tools within instructional coaching programs. Coaching can provide dedicated time and expert guidance, helping teachers contextualize the feedback metrics and translate them into actionable steps. These tools can also enhance coaching by providing detailed, time-stamped evidence that coaches and teachers can inspect together, much as they would analyze video or student work. And coaches can create social accountability structures that motivate teachers to engage with the feedback on a regular basis. In this integrated model, automated feedback becomes another instructional artifact that affords coaching conversations grounded in reasoning about classroom evidence.

Despite promising evidence that automated feedback can function as an instructional artifact for evidence-based coaching, important questions remain about feasibility, interpretation, and uptake. At minimum, such tools can help in coaching contexts where little classroom evidence is available, yet availability alone may not produce productive reflection. Skepticism

about accuracy, challenges of interpretation, and limited available metrics can impede their meaningful use. To better understand the value and limits of these tools, we examine coaches' experiences with M-Powering Teachers, asking what opportunities and challenges they anticipate, how they incorporate the evidence in real conversations, and how they perceive the tool after initial use.

The M-Powering Teachers Application

Automated teacher feedback tools tend to share a common workflow: Teachers upload audio, the system transcribes it, and computational analyses generate feedback on talk time and use of specific teacher and student talk moves. In this study, we use the M-Powering Teachers application (MPT; https://www.mpoweringteachers.com/), a freely available tool with specialized focus on math instruction.

What is Being Analyzed by MPT?

Users of MPT may upload either an audio file—automatically transcribed by MPT—or a pre-existing transcript. Transcription accuracy varies depending on the quality of the source audio. (To ensure consistency in transcription quality as much as possible, for this study we used a manual transcription service to obtain transcripts of classroom recordings.) MPT then computes teacher–student talk ratios and automatically tags key talk moves—for example, teachers' uptake of student ideas, use of focusing questions, and student reasoning—using classifiers trained on mathematics classroom corpora (Alic et al., 2022; Demszky & Hill, 2023; Demszky et al., 2021).

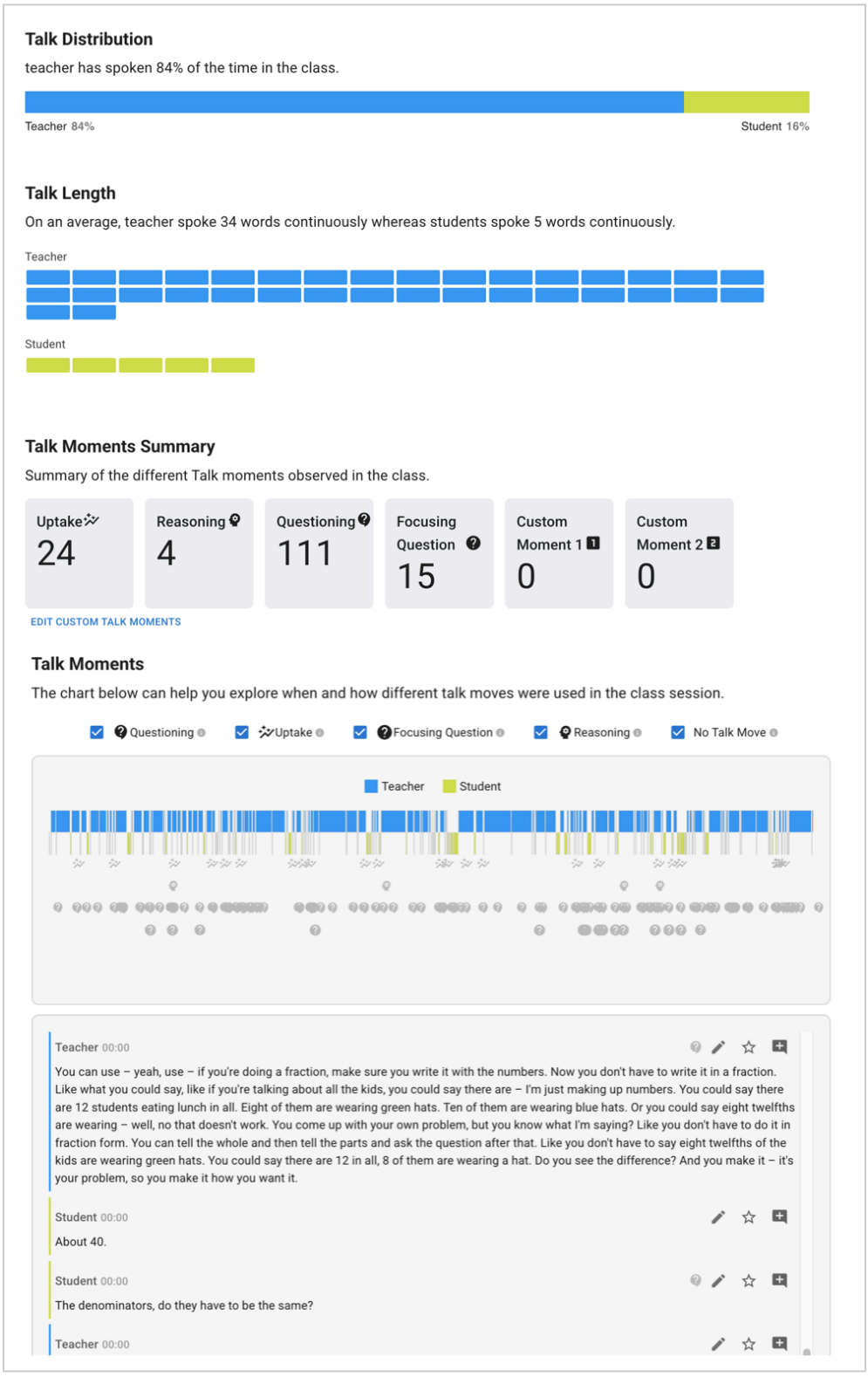
User Interface

When processing finishes, MPT displays a dashboard of discourse-based insights, as shown in Figure 1. These insights include visuals of the teacher–student talk-time ratio and the

mean number of words spoken by the teacher and by the students. The dashboard also shows a frequency count of different types of “talk moments” (i.e., talk moves) identified in the transcript, and a clickable transcript timeline that jumps to any selected moment. Teachers and coaches can mark noteworthy lines in the transcript, add comments, and correct or refine autogenerated labels. For example, they may manually add talk-move labels that were missed by the model or remove incorrectly labeled talk moves. Users may also create up to two custom talk-move categories, name them, and annotate exemplar transcript lines.

Figure 1

Screenshot of the MPT Feedback Dashboard



Methods

Study Participants

For our initial interviews, we recruited 20 instructional coaches through the authors' professional networks during the 2023–2024 school year. Most of these coaches supported mathematics teachers and most were affiliated with one of two U.S. university-based instructional coaching programs designed to support instructional quality—Program M and Program C [anonymized for peer review]. Among these initial interviewees, we invited six to continue participating in the study by piloting MPT in coaching sessions during the same school year. An additional four pilot coaches were recruited from Program M's professional network to pilot MPT in the next phase of the study.

These ten pilot coaches were selected to encourage diversity of perspectives toward using technology, to prioritize using well-defined coaching structures to support math teachers, and to select for those who had capacity and access to teachers to participate. Ultimately, one coach was removed from the pilot because they coached only a music and art teacher during the study time period. The remaining pilot coaches varied in years of coaching experience (range = 2–12+ years), geographic region (California, Minnesota, Texas, Colorado), grade level of coached teachers (K–12), and concerns and affordances about automated feedback expressed in initial interviews.

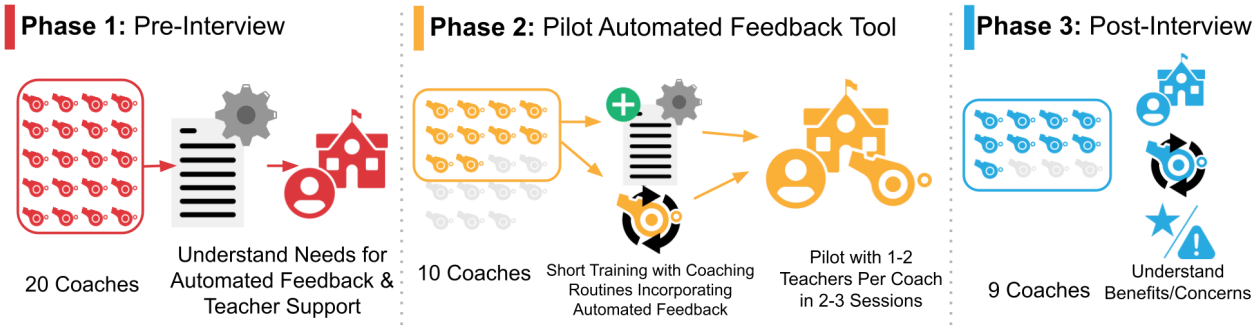
Procedure

The initial semi-structured interviews with 20 coaches focused on information about their role, their program of coaching, the teachers and students they serve, and the challenges they were facing. After being shown a prototype of MPT (see Figure 1), coaches were asked to share their concerns and imagine possibilities for the use of automated feedback tools (see full

interview protocol in Appendix A). In some cases, the researchers were able to make small changes to MPT based on coaches' suggestions (e.g., adding a disclaimer about accuracy, including definitions of each detected feature, allowing users to click directly on the transcript to automatically scroll to a specific moment).

Figure 2

Overview of Our Study Procedure



In the next phase of the study, we followed the subset of coaches as they piloted MPT. As part of the pilot onboarding process, they received training materials—a set of routines to use MPT aligned with their respective coaching programs—and attended two 1-hour orientation sessions on Zoom. The training materials and routines were co-designed by three of the authors and expert coaches from Program C and Program M, and the goal was to provide options and examples for how to use MPT to engage teachers in reflection about previous instruction and planning for new lessons.[1] After completing the onboarding, each pilot coach was video-recorded during two to three coaching sessions with one to two different teachers.

The first coaching session between each coach–teacher pair was recorded before they used MPT to test recording equipment and to ensure that each pair would have at least one meeting before piloting MPT. Our ability to use these initial coaching sessions for comparison

[1] These materials are available on OSF: https://osf.io/n5hr9/?view_only=bf24cfcac12b49648d8b7a5c55825fa6

was limited by the fact that some coach-teacher pairs used this first session to plan for an upcoming lesson observation and therefore did not have any previous lesson's evidence to analyze together.

After this first recorded session, coaches had the opportunity to utilize automated feedback in their subsequent recorded coaching sessions. Researchers created user accounts on MPT for teachers and coaches. To receive MPT feedback, the teacher recorded the audio of a lesson they taught using a unique Zoom link provided by the researchers. Researchers sent the audio for improved transcription to a human transcription service (transcribe.me) and then uploaded the transcript to MPT, sharing the resulting feedback with the teachers and coaches. The feedback was available to them within 1–7 days (given the turnaround time of manual transcription), and teachers could then share the feedback with their coaches.

After the pilot, we invited all nine remaining pilot coaches to participate in a reflective interview. The interview centered on coaches' experiences using MPT, the benefits they perceived, and their concerns (see protocol in Appendix B). Coaches were also able to give feedback on the coaching training materials and routines. Finally, once our analysis of the interview data was complete, seven of the nine pilot coaches responded to our request to validate our representations of their ideas through member-checking (Birt et al., 2016).

Our analytic sample included 20 initial interviews, 41 recorded coaching conversations of nine coaches working with 16 different teachers,[2] and an additional nine post-pilot reflective interviews. Each coaching conversation lasted, on average, about 32 minutes. Sixteen of the conversations were recorded before the coach and teacher pair utilized MPT; in 25, the coach and teacher had the opportunity to utilize feedback generated by MPT on a lesson the teacher taught

[2] We excluded one pilot coach's recordings due to her working with an art and a music teacher and MPT not being applicable to their coaching conversations.

since the previous coaching conversation.

Analysis

We leveraged initial interview data to identify anticipated affordances and limitations of MPT. All 20 of these interviews were open-coded by the research team (Given, 2008) into categories of affordances, concerns, and suggestions for minor improvements to MPT features. The research team discussed coding until consensus was reached and worked with a software engineer to implement as many minor suggested improvements to MPT as possible before coaches and teachers began piloting the tool. The coding procedure was similar for the nine post-pilot interviews, which we used to understand coaches' perceptions of the tool after they had used it in several coaching sessions.

Looking at our observation data, we conducted interaction analysis to understand how coaches utilized MPT in their conversations with teachers. We first sought to establish what kinds of moves coaches were using to better understand which, if any, were enacted with the use of MPT. We applied Boguslav's (2024) Systematic Framework for Describing Coaching Discourse in a comprehensive coding of coaching moves across all recordings. The same coach utterance could be coded with multiple moves. Each utterance was also coded for whether it utilized MPT, which we defined as any reference to specific MPT metrics, diagrams, or transcript excerpts while enacting a coaching move.

Over the course of 25 coaching conversations, totaling 13.5 hours, we observed coaches employing a total of 847 coaching discourse moves. About 40% of these moves (328 out of 847) involved the use of MPT. The discourse moves most frequently observed with the use of MPT were (a) making observations, (b) giving interpretations, and (c) asking for teacher interpretations. This analysis suggested that MPT was mainly being used as a source of evidence

for making observations and offering interpretations of teachers' instruction.

Our secondary analysis focused on categorizing the *ways* in which MPT was used to supply evidence in these coaching conversations. Three researchers independently coded the moves where MPT was utilized, wrote analytic memos, and discussed discrepancies until consensus was reached (Cascio et al., 2019). The discussion of analytic memos and triangulation with codes from coach interviews inductively led to the identification of four primary themes in how coaches used MPT: (a) demonstrating effective practices; (b) probing; (c) challenging; and (d) comparing and generalizing. Two other interesting cases—the use of custom annotation and utilizing tool inaccuracy—were also observed. These themes were then applied in a secondary round of coding, where another set of analytic memos was written and discussed (Engle et al., 2014).

Findings

Research Question 1: Coaches' Initial Views of Automated Feedback

During initial interviews, coaches anticipated several potential affordances and concerns related to using automated feedback tools—and specifically, MPT— in their practice. Table 2 summarizes key themes in those interviews. Many of the anticipated benefits centered on how automated feedback affords novel and more effective ways of producing and using evidence in coaching. More than half of the coaches noted the tool's ability to visually display patterns of evidence through diagrams such as teacher–student talk ratios or the phases in the flow of a lesson. For example:

> It's a really cool way to visualize certain components of a lesson that I think I would have had a hard time putting some kind of quantitative number to, like…some version of how students entered this lesson. How much were students' ideas driving the lesson? And I

think that the tool made it a lot easier to add some numbers to that.

Table 2

Anticipated Affordances and Concerns of 20 Coaches

Affordance	n/20	Concern	n/20
Visual representations of abstract patterns	15	General **inaccuracy**	9
Ability to **revisit specific moments** using transcript	13	**Privacy**	9
Objectivity and specificity of evidence	8	**Low detection of student talk**, especially in small groups	6
Saving time or creating **efficiency** for coaches and/or teachers	7	Feedback can be **overwhelming** for teachers	6
More teacher **access to coaching**	7	Wasting time or **creating unproductive extra work** for coaches and/or teachers	5
Encouraging more or **better teacher reflection**	5	Incentivizing **unintended consequences**	5
Entry point for teachers reluctant to be video recorded	3	Transcripts don't include **gestures and tone**	3
Alignment with other professional learning initiatives	3		

Over half of the coaches also pointed out that the tool allows teachers to revisit specific moments of instruction by linking metrics directly to the classroom transcript and grounding reflections in evidence of what was said. Eight coaches also reported expecting the type of "objective" evidence provided through automated feedback to be beneficial. Seven imagined that a tool like this could save coaches time by eliminating the need to scribe or collect low-inference evidence during observations. For example, one coach shared the potential for the tool to help curate moments of interest to review with teachers because "no one really wants to comb through an hour's worth of video footage and hand scribe something like 'Oh, at Minute 38, a student

asked a really powerful question.'" Five coaches hypothesized that the evidence produced by MPT would enable more and better teacher reflection. In addition to evidence-related affordances, coaches noted that MPT might expand teachers' access to coaching by scaling coach reach, providing a lower-barrier entry point via audio-only recording for those teachers reluctant to be video-recorded, and align with broader existing professional learning initiatives (e.g., increasing student talk time).

Coaches also shared concerns about automated feedback before testing out the tool. Evidence emerged as a common thread among concerns as well, particularly regarding whether the available evidence would be sufficient (e.g., low detection of student talk or exclusion of gestures and tones) or trustworthy (e.g., inaccuracies in both transcripts and metrics). Some coaches pointed out that MPT would only work during whole-class discussions or in a participation structure with only one speaker at a time. They wondered about the accuracy of transcriptions for lessons in which students work in small groups or pairs and about the recording equipment that would be necessary to capture those moments. Some coaches wondered about potential structural inaccuracies, or biases, embedded in automated feedback. For example:

> One other question I have is about the AI coding these…things, and like…Who is coding that? Who is deciding what counts as reasoning or uptake, or even asking a question? And where that ends up might depend on the coders.

Apart from accuracy, coaches raised process-level concerns about using MPT, echoing challenges common to other recording-based approaches. Nine coaches noted the potential of recordings to infringe on teacher and student privacy. Some coaches shared that using automated feedback might result in incentivizing unintended consequences for teachers and students such as narrowing instruction to artificially inflate a particular metric. Five coaches also wondered if

automated feedback tools would create unnecessary extra work for teachers, such as learning how to use them. One coach related an attempt to introduce video-based coaching prior to this study: “But what ended up happening was, they felt like [reviewing video] was like a homework assignment. They already had enough on their plate.”

The prevalence of the theme of evidence in the interviews, which corroborated the intended purpose of the tool, led to us choosing evidence as a primary lens of analysis for the pilot.

Research Question 2: Coaches’ Use of Automated Feedback

We now turn to answering how coaches utilized automated feedback in their coaching conversations. Specifically, we discuss the results of our secondary round of coding, focusing on the ways in which coaches used MPT as evidence by (a) demonstrating effective practices; (b) probing; (c) challenging; and (d) comparing and generalizing. We also elaborate on the two other coded themes, the use of custom annotation and utilizing tool inaccuracy.

In some cases, coaches used MPT to demonstrate effective practices, as seen in comments like, “So here’s a good uptake one” or “Here we see you ask a regular question, a focusing question, and that there’s also some uptake.…So that’s a moment where, like, you’ve taken what she said and are sort of building on that idea.” Coaches used these examples to encourage teachers to do more of a similar move, make connections to a shared instructional framework, or demonstrate connections between teacher and student talk.

In other instances, coaches used detected moments to probe for elaboration, asking teachers to add details or justification: “You asked a question here, so tell me more about what is happening here,” or “And then what do you see happen because of that question?” These probes

elicited more information about teacher intent or decision making. Sometimes probes directed teachers to notice patterns of cause and effect between their own and students' talk.

At times, coaches used evidence from MPT to challenge or push back on teacher claims or observed practice. These moments often involved constructive questioning aimed at refining the form or intent of a specific talk move (e.g., "Could you rephrase this [question that we just read you asking in the transcript] so that it's not a yes or no?"; "How can we get the follow-up to give away less but still get them to go where you want them to go?"; "Are there other places where you thought 'I could have maybe had a moment of asking questions here'?"). Challenges prompted teachers to revisit moments of instruction to reimagine decisions they had made and how similar moves might play out in future lessons.

Some coaches helped teachers make comparisons or draw generalizations across detected moments. In one example, a coach asked a teacher to examine three separate examples of student reasoning identified by MPT and justify why they should be categorized that way or not. Another coach prompted reflection by contrasting the effects of different talk moves: "Let's look at all the ones that did lead to reasoning and see what they have in common, and let's look at the ones that didn't lead to reasoning and see why." Additionally, several coaches used MPT's talk distribution visualization to analyze the balance of teacher and student talk, or they discussed mapping transcript segments to instructional phases of the lesson, such as whole-class discussions or independent student work. These comparisons and generalizations encouraged teachers to step back from a focus on individual moments to navigate broader patterns of interactions or transcript segments.

Another strategy involved custom annotation, whereby coaches tagged specific portions of the transcript to capture additional interactional features not automatically identified by MPT.

This emerged as a powerful tool for tailoring evidence-based discussions. In some cases, coaches annotated phrases such as “How do you know?” to identify talk moves like “eliciting student ideas” not included in MPT’s default metrics. In other cases, coaches disaggregated detected focusing questions into subcategories—such as “why,” “what,” or “how”—to examine how different question types elicited varied student responses. Custom annotations enabled coaches to build on MPT’s current capabilities for automatic detection by including other relevant features of classroom interactions.

We observed a notable case involving some of the unexpected ways in which coaches and teachers engaged with the perceived inaccuracy of MPT. Throughout the coaching sessions, we documented multiple instances in which coaches disagreed with or questioned transcription or automatically detected moves. Interestingly, despite these concerns, we observed instances of coaches leveraging the tool’s inaccuracies as a pedagogical opportunity. For instance, several prompted teachers to evaluate the veracity of particular metrics, asking questions like “So it identified…that it’s a question.…That type of question is a regular question, not a focus question.…Does that feel right?” or “Now with some of that context, talk to me a little bit more about how you see that [specific student talk] as reasoning.” These moves encouraged teachers to think about what qualifies as a focusing question or student reasoning in the context of their classroom interactions. Other coaches encouraged teachers to identify false negatives by asking, “Are there other moments that you notice…the tool missed?”

Coaches also explored possible explanations for over- or under-identification of talk features, as in the following reflection: “What was…jumping out to you in terms of what was contributing to the average student talk length and to your own talk length? Because my experience, when I was there in the classroom, was that you were very minimally talking.” These

interactions suggest that, in some instances, coaches accepted the tool's limitations as a springboard for reasoning about the constructs MPT aimed to capture. One coach, during a coaching conversation with a teacher, even speculated that discrepancies between teacher and coach perceptions of MPT's metrics might itself catalyze reflection:

> Yes, I completely agree that it's not quite picking up everything, and I wonder if sometimes seeing those numbers—the dissonance that occurs for us when we see, like, "Oh, here's the actual numbers compared to what it felt like" might provoke deeper thinking.

Another coach proposed that simply knowing MPT was "keeping track" might have been enough to prompt behavioral changes: "Do you feel like it helped you to ask more questions just by thinking about it?... They're keeping track." In these ways, coaches transformed the limitations of MPT into generative moments for instructional reflection and growth.

Research Question 3: Coaches' Reflections After Piloting Automated Feedback

After piloting MPT, all but one of the nine coaches reaffirmed benefits relating to the use of evidence in coaching conversations. Some reported teachers' reflections as more factually grounded (e.g., "I think it'll be a really good tool that teachers can reference back to and do self reflection on…especially if you're hearing yourself.…[MPT] took exactly what you were saying."). Others described how the tool enabled new kinds of teacher reflection that would be difficult to achieve without evidence from automated feedback. For example, this coach described how MPT enabled teachers to set goals they wouldn't necessarily otherwise consider:

> I don't think the teacher would have set that goal without knowing that it was something that could be measured by the tool. I don't think they would have been like, "All right. I want you to sit there and count how many words the student is saying continuously

> versus me."

These insights corroborate the anticipated benefits expressed in initial interviews and interactions observed in coaching conversations.

Half of the piloting coaches elaborated on the potential of automated feedback to save them and the teachers time by creating efficient ways to curate evidence of particular moments to spark conversations: "I think where the time is saved is the [automatic detection of talk moves]…without that…things run a little bit slower because I'm having to guide them to [specific moments I want them to notice]." Automatic detection of talk moves sometimes helped narrow the amount of video or transcript a coach needed to look through to find relevant examples. However, two pilot coaches reported MPT not saving as much time as they expected it to or creating other work to fill that time (e.g., "It just minimizes…the time it takes to make notes and reflect on the notes. You just reflect on what is recorded, which is still work, right?"). Time saved in documenting evidence of what occurred in a classroom might have been offset by time spent interpreting that evidence. Regardless, it is important to note that each pilot coach only used MPT a total of two to four times, and thus its potential efficiency in curating evidence might not yet have been fully realized.

A concern expressed by almost every piloting coach was the tool's inability to accurately capture evidence of student talk, especially in small groups. As one coach explained,

> But then that also meant if we ever had a goal around, what kind of student talk we wanted —so we want students to be asking each other questions, or we want them to be explaining—we couldn't get it, because we just didn't have that data.

Thus, the types and amounts of evidence provided by MPT constrained the types of reflective conversations that could be based on them.

Importantly, half of the pilot coaches expressed concerns that the evidence presented by the tool could be overwhelming and/or unwelcoming for teachers. One coach shared,

> I could see a coach who maybe isn't very skilled or very practiced…looking at this data with a teacher, and it becoming very targeted, of every moment that's going wrong, and "here are all of the things that you should do differently," and like just all of those sort of negative opportunities to really sort of disempower teachers. And that it could potentially take a lot of skill, both for teachers and for coaches, to be able to reflect in ways that are helpful and authentic.

Indeed, several coaches spoke to the importance of framing and presentation, both in terms of the tool's interface and the facilitation by a coach, in making the evidence actionable for teachers.

Finally, the majority of pilot coaches expressed favorable views of what they perceived as the objectivity of the data presented by MPT. This perception may support the building of rapport between coaches and teachers through coaches' framing of MPT as an external observer—regardless of whether it should be considered truly objective.

Discussion

Across interviews and coaching session analyses, we observed ways in which automated feedback was reported and utilized as an instructional artifact. Some of the most salient themes and controversies with respect to its usage included the role of evidence in discussions between coaches and teachers. Some MPT pilot coaches had not previously used video in their coaching, and a commonly cited reason was reluctance from teachers or administrators to allow classroom recording due to privacy concerns. For coaches like these, MPT may offer an alternative means of accessing some of the evidence-based insights that video provides while also supporting student privacy through features such as transcript de-identification. For coaches already using

video, an important question is what additional value automated feedback might bring beyond the evidence already captured through recorded instruction.

Most coaches shared that they like the "objectivity" of the MPT metrics. When deemed accurate by coaches, we observed coaches using them in a variety of ways to ground discussions in the evidence they provided. Whether or not the information provided by MPT should be considered objective, we saw a potential advantage in how coaches and teachers positioned it as "another pair of eyes." One coach spoke specifically to the benefits of MPT's objectivity: "When *you* [emphasis added] show a teacher, that is very helpful compared to me saying what I've seen, cause it's actually proving this is what came out of [the teacher's] mouth." A particular aspect of this insight is particularly worth highlighting: MPT can be positioned as an additional vantage point in coaching conversations. This coach spoke to the helpfulness of "you" (MPT) presenting information about the teacher's instruction. This positioning may have implications for grounding reflections in evidence and reasoning about potential inaccuracies, but also for rapport between coaches and teachers. When *MPT* shows a teacher some information, the coach and teacher can decide how to respond together.

Notably, however, coaches did not always endorse MPT's metrics as accurate representations of what happened in classrooms. This was sometimes due to classroom talk, particularly student talk, that was not captured in transcription. Some coaches noted that, regardless of how much talk was transcribed, transcripts excluded information about gestures, tone, and gaze that could help contextualize important aspects of lesson participation. Additionally, many coaches pointed to inaccuracies in talk-move detection, even for sections of lessons that were transcribed. We, as authors, agree that the tool, especially in this stage of its development, has many errors. As some coaches pointed out, these errors could be demoralizing

for teachers. However, we observed instances of coaches taking advantage of MPT as an imperfect observer to facilitate teacher reflection by asking them to reason about whether they should agree with MPT's conclusions. As MPT and the underlying technologies that make it possible (e.g., automatic speech transcription) continue to improve, the amount of available information about classroom interactions is likely to increase. There will, however, continue to be some amount of error as automated feedback tools miss and misidentify classroom talk. We hope that these errors can continue to be facilitated in ways that allow for curiosity, reflection, and professional learning.

When coaches trusted the accuracy of MPT evidence (and in some cases when they didn't), we have seen how it can encourage teachers and coaches alike to focus on particularly high-leverage aspects of classroom observations, such as specifics of student thinking or participation in learning. Not all instructional artifacts used in coaching afford these possibilities. In one contrasting example, we observed a coach–teacher pair examining a dashboard from a commonly used benchmark assessment that listed students' names next to the percentage of questions they answered correctly in an assessment on representing story problems with equations. The coach asked the teacher to interpret the evidence provided by the dashboard: "Would you say [these students] are getting it or not?" The teacher was asked to sort the students into a graphic organizer with columns for "Does Not Meet [Expectations]," "Approaches [Expectations]," "Meets [Expectations]," and "Exceeds [Expectations]" based on the dashboard-provided evidence (i.e., percentage of questions answered correctly). When the conversation turned to potential factors contributing to the observed student results, the teacher and coach discussed "testing stamina."

We share this example for two reasons: One, the discussion of testing stamina implies

that many students were, in the opinion of the coach–teacher pair, being overtested. This could indicate a need for other sources of evidence about students' learning (e.g., talk measured by automated feedback tools) that could supplement or replace some of the benchmark assessments they were taking. Another reason is to highlight that the types of student evidence we examine afford different possibilities for interpretation and for responding. In this example, no information was preserved about why students answered questions incorrectly or how those students were thinking about the topic of representing written contexts with equations. While the evidence provided allows for interpretations of who "got it" or not and for responses such as sorting students into proficiency levels, it does not enable responses like designing a timely question to ask a specific student to confront a particular misconception. These latter kinds of pedagogical responses require sources of evidence that are closer to how students are thinking about mathematics.

Limitations and Future Work

While the coaches' perspectives and the interactions we observed show some possibilities for MPT's use, they should not be generalized from such a small sample. Although most of the core functionalities of MPT (e.g., detection of talk dynamics and teacher moves) are shared by other automated feedback tools like TeachFX and TalkMoves, the insights observed for MPT may not directly generalize to those tools. For example, the user interface and the parts of the transcript highlighted in the application may have implications for coach usability. Additionally, each coach used MPT just two to four times during this study. Longer-term use of the tool would likely reveal other insights.

To better realize the potential of MPT and similar tools as instructional artifacts for evidence-based teacher reasoning, we need additional reliable measures of student talk moves.

These metrics would ideally enable teacher reflection upon "individual students' knowledge, ideas, and intentions" (Cohen & Ball, 1999, p. 10), which is one of the key factors in making instructional artifacts useful for professional learning (Ghousseini & Sleep, 2011). Teacher reflection on artifacts has been shown to be especially valuable when directed toward student thinking and interactions (Anantharajan, 2020; Sherin & van Es, 2005; Stockero et al., 2017). Developing these measures will require improving the capture and transcription of student talk across participation structures (whole-class, small-group, pairs). These metrics would enable us to make connections between specific instructional practices and their impacts on students (Amador et al., 2025).

Beyond the characteristics of individual coaches, teachers, and their activities, effective use of MPT within coaching sessions depends on system-level conditions (Cobb et al., 2020). These include sustained funding and administrative backing; leaders' understanding of targeted instructional practices and change management; professional learning for leaders on cultivating coaching-supportive climates; and equitable, transparent hiring that yields coaches respected by principals and teachers (Bengo, 2016). Power relations among teachers, coaches, and other stakeholders also shape what counts as evidence and whose interpretations prevail (Lefstein & Snell, 2011). Accordingly, future work should examine how automated feedback tools are implemented under varied conditions, how to design aligned professional learning for teachers and coaches, and how these interventions interact with concurrent district reforms.

A parallel priority is to study the complementary roles of multiple artifacts—automated feedback, curated video clips, student work samples, and observation notes—and to identify effective combinations and sequences for different contexts (e.g., screening with automated metrics to target subsequent video analysis). This line of work should test dosage and timing by

grade level, content area, and teacher experience; assess cost-effectiveness; and develop responsible approaches to aggregating data for improvement while avoiding perverse incentives.

Conclusion

In this study, we examined how instructional coaches anticipated, enacted, and reflected on automated feedback as an instructional artifact. Our findings suggest that automated tools like MPT can productively anchor coaching conversations in classroom evidence by making patterns of teacher–student discourse more visible and accessible. Coaches used MPT to highlight effective moves, probe and challenge teacher reasoning, and generalize across moments of instruction—at times even leveraging inaccuracies as opportunities for reflection. At the same time, limits in transcription accuracy and capture of student talk mean that productive use depends on skilled facilitation, alignment with context, and continued technical improvement. Overall, automated feedback should be positioned as a complementary artifact—alongside video, student work, and observation notes—rather than a standalone solution. When thoughtfully integrated into existing professional learning routines, it can expand evidentiary resources and deepen evidence-based reasoning in support of teacher learning and student understanding.

References

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesh (Eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 224–233). Association for Computational Linguistics. https://aclanthology.org/2022.bea-1.27/

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness*, 8(4), 475–489. https://doi.org/10.1080/19345747.2015.1017680

Amador, J. M., Gillespie, R., & Carson, C. (2025). Coaches and teachers shift noticing across coaching cycles: Analysis of video-based annotations. *Research in Mathematics Education*, *27*(1), 1–22. https://doi.org/10.1080/14794802.2023.2212258

Anantharajan, M. (2020). Teacher noticing of mathematical thinking in young children's representations of counting. *Journal for Research in Mathematics Education*, *51*(3), 268–300. https://doi.org/10.5951/jresemtheduc-2019-0068

Bengo, P. (2016). Secondary mathematics coaching: The components of effective mathematics coaching and implications. *Teaching and Teacher Education*, *60*, 88–96. https://doi.org/10.1016/j.tate.2016.07.027

Birt, L., Scott, S., Cavers, D., Campbell, C., & Walter, F. (2016). Member checking: A tool to enhance trustworthiness or merely a nod to validation? *Qualitative Health Research*, *26*(13), 1802–1811. https://doi.org/10.1177/1049732316654870

Boguslav, A. (2024). Parsing coaching practice: A systematic framework for describing coaching discourse. *AERA Open*, *10*. https://doi.org/10.1177/23328584241263861

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education*, *24*(2), 417–436. https://doi.org/10.1016/j.tate.2006.11.012

Brophy, J. (Ed.). (2003). *Advances in Research on Teaching, Vol. 10. Using video in teacher education*. Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-3687(03)10012-0

Calandra, B., Gurvitch, R., & Lund, J. (2008). An exploratory study of digital video editing as a tool for teacher preparation. *Journal of Technology and Teacher Education*, *16*(2), 137–153. https://www.learntechlib.org/primary/p/23496/

Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*(3), 430–454. https://doi.org/10.1086/657654

Cascio, M. A., Lee, E., Vaudrin, N., & Freedman, D. A. (2019). A team-based approach to open coding: Considerations for creating intercoder consensus. *Field Methods*, *31*(2), 116–130. https://doi.org/10.1177/1525822X19838237

Center for Education Policy. (2023). *MQI coaching cycle.* https://mqicoaching.cepr.harvard.edu/mqi-coaching-cycle

Center to Support Excellence in Teaching. (2021). *Our approach*. https://cset.stanford.edu/about/our-approach

Chen, G., Chan, C. K. K., Chan, K. K. H., Clarke, S. N., & Resnick, L. B. (2020). Efficacy of video-based teacher professional development for increasing classroom discourse and student learning. *Journal of the Learning Sciences*, *29*(4–5), 642–680. https://doi.org/10.1080/10508406.2020.1783269

Chen, G., Clarke, S. N., & Resnick, L. B. (2015). Classroom discourse analyzer (CDA): A discourse analytic tool for teachers. *Technology, Instruction, Cognition and Learning*, *10*(2), 85–105.

Cobb, P., Jackson, K., Henrick, E., & Smith, T. M. (2020). *Systems for instructional improvement: Creating coherence from the classroom to the district office*. Harvard Education Press.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement* (CPRE Research Report No. RR-43). Consortium for Policy Research in Education. https://www.cpre.org/sites/default/files/researchreport/783_rr43.pdf

Coles, A. (2014). Mathematics teachers learning with video: The role, for the didactician, of a heightened listening. *ZDM: The International Journal on Mathematics Education*, *46*(2), 267–278. https://doi.org/10.1007/s11858-013-0541-3

Demszky, D., & Hill, H. (2023). The NCTE transcripts: A dataset of elementary math classroom transcripts. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesh (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 528–538). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.44

Demszky, D., & Liu, J. (2023). M-Powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes. In *Proceedings of the Tenth ACM Conference on Learning@ Scale* (pp. 59–69). https://doi.org/10.1145/3573051.3593379

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 483–505. https://doi.org/10.3102/01623737231169270

Demszky, D., Liu, J., Hill, H. C., Sanghi, S., & Chung, A. (2025). Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further

enhancement. *Computers & Education*, *227*, Article 105183. https://doi.org/10.1016/j.compedu.2024.105183

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. B. (2021, August). Measuring conversational uptake: A case study on student–teacher interactions. In C. Zozng, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1638–1653). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.130

Desimone, L. M., & Pak, K. (2017). Instructional coaching as high-quality professional development. *Theory Into Practice*, *56*(1), 3–12. https://doi.org/10.1080/00405841.2016.1241947

Engle, R. A., Conant, F. R., & Greeno, J. G. (2014). Progressive refinement of hypotheses in video-supported research. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 239–254). Routledge. https://doi.org/10.4324/9780203877258

Erickson, F. (2014). Ways of seeing video: Toward a phenomenology of viewing minimally edited footage. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 145–155). Routledge. https://doi.org/10.4324/9780203877258

Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review*, *16*, 41–67. https://doi.org/10.1016/j.edurev.2015.06.001

Ghousseini, H., & Sleep, L. (2011). Making practice studyable. *ZDM: The International Journal on Mathematics Education*, *43*(1), 147–160. http://doi.org/10.1007/s11858-010-0280-7

Given, L. M. (Ed.). (2008). *The Sage encyclopedia of qualitative research methods*. Sage publications.

Hmelo-Silver, C. E., Chernobilsky, E., & Jordan, R. (2008). Understanding collaborative learning processes in new learning environments. *Instructional Science*, *36*(5), 409–430. https://doi.org/10.1007/s11251-008-9063-8

Horn, I. S., Garner, B., Kane, B. D., & Brasel, J. (2017). A taxonomy of instructional learning opportunities in teachers' workgroup conversations. *Journal of Teacher Education*, *68*(1), 41–54. https://doi.org/10.1177/0022487116676315

Jacobs, J., Scornavacco, K., Clevenger, C., Suresh, A., & Sumner, T. (2024). Automated feedback on discourse moves: Teachers' perceived utility of a professional learning tool. *Educational Technology Research and Development*, *72*, 1307–1329. https://doi.org/10.1007/s11423-023-10338-6

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*, Article 103631. https://doi.org/10.1016/j.tate.2022.103631

Kochmanski, N., & Cobb, P. (2023). Identifying productive one-on-one coaching practices. *Teaching and Teacher Education*, *131*(2), Article 104188. http://dx.doi.org/10.1016/j.tate.2023.104188

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Lefstein, A., Louie, N., Segal, A., & Becher, A. (2020). Taking stock of research on teacher collaborative discourse: Theory and method in a nascent field. *Teaching and Teacher Education*, *88*, Article 102954. https://doi.org/10.1016/j.tate.2019.102954

Lefstein, A., & Snell, J. (2011). Professional vision and the politics of teacher learning. *Teaching and Teacher Education*, *27*(3), 505–514. https://doi.org/10.1016/j.tate.2010.10.004

Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record*, *91*(4), 509–536. https://doi.org/10.1177/016146819009100403

Oshima, J., Oshima, R. & Matsuzawa, Y. (2012). Knowledge Building Discourse Explorer: A social network analysis application for knowledge building discourse. *Educational Technology Research and Development, 60*, 903–921. https://doi.org/10.1007/s11423-012-9265-2

Prusak, K., Dye, B., Graham, C., & Graser, S. (2010). Reliability of pre-service physical education teachers' coding of teaching videos using Studiocode analysis software. *Journal of Technology and Teacher Education*, *18*(1), 131–159. https://www.learntechlib.org/primary/p/29396/

Roschelle, J., & Goldman, S. (1991). VideoNoter: A productivity tool for video data analysis. *Behavior Research Methods, Instruments, & Computers*, *23*(2), 219–224. https://doi.org/10.3758/BF03203368

Russell, J. L., Correnti, R., Stein, M. K., Thomas, A., Bill, V., & Speranzo, L. (2020). Mathematics coaching for conceptual understanding: Promising evidence regarding the Tennessee math coaching model. *Educational Evaluation and Policy Analysis*, *42*(3), 439–466. https://doi.org/10.3102/0162373720940699

Sherin, M., & van Es, E. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education*, *13*(3), 475–491. https://www.learntechlib.org/primary/p/4824/

Stockero, S. L., Rupnow, R. L., & Pascoe, A. E. (2017). Learning to notice important student mathematical thinking in complex classroom interactions. *Teaching and Teacher Education*, *63*(2), 384–395. https://doi.org/10.1016/j.tate.2017.01.006

Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., & Sumner, T. (2021). *Using transformers to provide teachers with personalized feedback on their classroom discourse: The TalkMoves application*. arXiv. https://doi.org/10.48550/arXiv.2105.07949

van Es, E. A., & Sherin, M. G. (2010). The influence of video clubs on teachers' thinking and practice. *Journal of Mathematics Teacher Education*, *13*, 155–176. https://doi.org/10.1007/s10857-009-9130-3

van Vondel, S., Steenbeek, H., van Dijk, M., & van Geert, P. (2018). The effects of video feedback coaching for teachers on scientific knowledge of primary students. *Research in Science Education*, *48*(2), 301–324. https://doi.org/10.1007/s11165-016-9569-z

Appendix A

Pre Interview Protocol

PART 1: Background (10 min)

1. Please tell us a little bit about yourself—your professional role, and the teachers and students you serve.
    a. *Probing: Tell us how you became a coach.*

PART II: Current Coaching Experiences (20 min)

2. We want to know about you as a coach and about your coaching role.
    a. Can you tell me how you work 1:1? (More specifically: Can you tell us what a typical coaching cycle looks like?)
        i. *Probing: Is there a particular coaching approach or coaching tools/resources at the site or district level that guide your work?*
    b. Do you currently use video, audio, or any other technology to support your coaching work? If so, what? How?
        i. *Probing: What coaching tools do you use to support your coaching?*
    c. How many teachers do you coach 1:1? How is that determined?
        i. *Probing: Time constraints, teachers interests, etc.*
    d. What does a typical day and week look like for you in your current coaching role?
3. What do you do well as a coach?
4. What aspects of your current coaching role do you find most frustrating or challenging?
    a. *Probing: Why?*
5. What are a few things you would change about your current role that you think would make you a more effective coach?
    a. *Probing: How would these changes help?*

PART III: Scenarios and Reactions (25 min)

Framing: Technology is transforming everyone's job, and we want to hear from coaches about how you feel technology might be best used in your role. Technology should never replace coaches, and we understand the value your relationship with your teachers has on your ability to coach. We're curious to understand your ideas around a platform that uses an automated feedback tool to analyze teacher transcripts or recordings. Teachers record a lesson, upload it to a platform, and get feedback on aspects of classroom dialogue—for instance, the amount of teacher versus student talk time, teachers' use of open-ended questions, or teacher "uptake" of student ideas. Teachers receive information about the frequency of use of these talk moves, and also see examples of how they used them with students. This feedback is nonevaluative and, because it is generated automatically, more objective than typical classroom observations.

From previous studies, we found that teachers had difficulty translating descriptive automated feedback into actionable improvements to their practice when they did so in isolation, so we'd

like to consider ways to integrate these tools in complementary ways to how coaches are already working with teachers.

For the next part of our interview, I am going to show you an example of data (not real data!) from a platform that analyzes teachers' transcripts or a classroom recording with analysis using an automated feedback tool. I will then ask you to take some time looking at sample feedback generated from the data in order to respond to a few questions. Do you have any questions before we begin?

6. Here are some samples of automated analysis and feedback the platform produced from the recording you took of Teacher A's class. You can see that this teacher is getting feedback on uptake, reasoning, and questioning. Take a moment to look at these samples. (*Give participant 3-4 minutes to explore and ask questions.)*
    a. What questions do you have about the analysis and feedback examples?
    b. What do you like about what you see? Why?
    c. What do you find less useful? Why?
    d. Would this data, in combination with your own notes from your observation of this class, be helpful for a coaching conversation with the teacher? Explain.
        i. *Probing: Is there anything missing that you think would be helpful to include?*
7. Now that you have seen examples of automated feedback:
    a. What barriers do you see in using this platform?
    b. Can you imagine a teacher you coach recording themselves and uploading it to the app for you to view?
    c. Are there other ways you might use the recordings and data as part of your coaching work?
        i. *Probing: In 1:1 coaching scenarios, in PLC team meetings, for understanding teachers' instruction and plan for professional development*

PART IV: Conclusion (5 min)

8. Do you have any other thoughts to share?
9. Do you have any questions for me?
10. Once this platform is developed, would you be interested in piloting this platform with teachers?
    a. Do you have colleagues who might be interested?

Appendix B
Post Interview Protocol

Interview Protocol (~60 min.)

Hi. Thank you for piloting our automated feedback tools to see how they might support your role as a coach and mitigate the most common challenges you face in coaching. We really appreciate your time and your support of your participating teachers!

My name is [Name], and I am part of the team of researchers from Stanford, Harvard, and the University of Maryland studying how automated feedback could potentially be integrated into teacher coaching, within existing processes and in ways that complement the core competencies and practices of a coach. I will be asking you some questions to understand your successes and challenges during the piloting of the feedback tools. I will also solicit any suggestions you have for adjustment to the tools and/or coaching protocols.

PART I: General Impressions of the Tool (15 min)

1. Please tell us about your experiences using the M-Powering Teachers automated feedback tool during your coaching conversations.
    a. What did you like about the tool?
        i. What did you like about that feature?
        ii. In what ways was it useful?
    b. What challenges did you encounter with the tool?
        i. Do you have any suggested improvements to overcome these challenges?
    c. What, if anything, surprised you about using the tool?

PART II: Supports, Scaffolds, and Coaching Protocols (15 min)

2. When we last spoke, we shared some examples of protocols (protocol 1, protocol 2) that could be used to incorporate automated feedback into coaching conversations. Did you use any of these protocols? If so, how?
    a. What did you like about the protocol(s) you used?
    b. What challenges did you encounter with the protocol(s)?
    c. What changes did you make to the protocol(s)? Why?

3. Are there other ways that you incorporated the automated feedback into your coaching conversations with teachers?
    a. How did this align (or not) with the coaching framework you use?

PART III: Specific Goals and Challenges (25 min)

4. Think back to one of the coaching conversations you had with teachers during this pilot. What was one goal you had for the teacher in that conversation? In what ways, if any, did the automated feedback tool or coaching protocol help support that goal?
    a. In what ways, if any, did the automated feedback tool or coaching protocol help teachers reflect back on their previous instruction?
    b. In what ways, if any, did the automated feedback tool or coaching protocol help teachers plan ahead for future instruction?
5. Overall, do you think that the automated feedback tool or coaching protocol "worked" with the teachers you coached? Why or why not?
    a. Could you see using the automated feedback tool or coaching protocol with all teachers, or are there some who it wouldn't work as well for? Why or why not?
6. [In the pre-interview, you mentioned some aspects of your current coaching you find frustrating or challenging. In what ways, if any, did the automated feedback tool or coaching protocol help mitigate these challenges?] or [What coaching challenges, if any, did the automated feedback tool or coaching protocol help mitigate?]
    a. Are there any other ways the automated feedback tool or coaching protocol might help mitigate these challenges?
    b. Did the use of the automated feedback tool or coaching protocol save you any time? If so, how?
7. What advice would you give another coach who was interested in incorporating an automated feedback tool like this into their coaching conversations?
    a. Do you think the use of this tool is sustainable for instructional coaches? If no, why not?

PART IV Looking Ahead (5 min)

8. Would you be interested in using M-Powering Teachers more in the future? If so, in what ways?
    a. Would you be interested in participating in future studies?

9. Do you have any other thoughts to share?
10. Do you have any questions for me?