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 applying an automated feedback tool to classroom recordings in order to support evidence-based reflection in coaching. We conducted semi-structured interviews with 9 coaches and coded 25 coaching conversations for coaching moves. Our analysis suggests that coaches used generated evidence to demonstrate effective practices, probe, challenge, and compare/generalize across moments. Coaches valued the tool's efficiency, visual representations, and perceived objectivity but raised concerns about accuracy, particularly for student talk. Coaching moves involving inferences or observations about instruction were significantly associated with utilization of automated feedback. 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: November 2026

Suggested citation: Malamut, James, Dorottya Demszky, Christine Bywater, Michele Reinhart, and Heather C. Hill. (2026). 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[1] (https://orcid.org/0009-0008-6618-5913), Dorottya Demszky[1] (https://orcid.org/0000-0002-6759-9367), Christine Bywater[1], Michele Reinhart[1], Heather Hill[2] (https://orcid.org/0000-0001-5181-1573)

1. Stanford University, Stanford, CA, USA
2. Harvard University, Cambridge, MA, USA

Author Note

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, Samantha Booth, Claire Gogolen, Jing Liu, Hannah Rosenstein, and the coaches and teachers who piloted this tool.

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

Correspondence concerning this article should be addressed to James Malamut, Stanford University, 485 Lausen Mall, Stanford, CA 94305. Phone: (650) 723-02109; Email: jmalamut@stanford.edu

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 applying an automated feedback tool to classroom recordings in order to support evidence-based reflection in coaching. We conducted semi-structured interviews with 9 coaches and coded 25 coaching conversations for coaching moves. Our analysis suggests that coaches used generated evidence to demonstrate effective practices, probe, challenge, and compare/generalize across moments. Coaches valued the tool's efficiency, visual representations, and perceived objectivity but raised concerns about accuracy, particularly for student talk. Coaching moves involving inferences or observations about instruction were significantly associated with utilization of automated feedback. 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.

Keywords: instructional coaching, evidence-based reflection, automated feedback, AI

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

Robust empirical evidence suggests that programs that contain 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 classroom reflection is often underutilized in coach–teacher discussions due to competing demands such as sourcing classroom materials, planning future instruction, or preparing for student testing (Gibbons & Cobb, 2017).

Classroom recordings are common coaching artifacts that can productively anchor evidence-based reflection (Amador et al., 2025; Gaudin & Chaliès, 2015). However, preparing such artifacts for coaching—recording, reviewing, clipping, and aligning to goals—is time-intensive and resource-demanding, which limits routine use. Tools that provide automated feedback based on classroom recordings 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. New advances in technology that enable automated feedback, such as automatic speech recognition and artificial intelligence (AI) language models, are still in their infancy. However, studying automated feedback tools, even with known limitations, can

potentially guide their development in ways that maximize the potential impact on teaching and learning. To examine the potential utility of such tools in evidence-driven coaching, we observed and interviewed nine mathematics instructional coaches as they conducted coaching conversations with M-Powering Teachers (MPT), an automated feedback tool. We asked:

1. What coaching moves were most likely to co-occur with the use of automated feedback?
2. How did coaches utilize automated feedback to share and elicit evidence-based reflection?
3. What affordances and limitations did coaches report after piloting an automated feedback tool?

Literature Review

Evidence-Based Reflection in 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 reflection upon observed instruction, may be especially productive (Kochmanski & Cobb, 2023). Evidence-based reflection includes describing, analyzing, or reasoning about classroom evidence and allows teachers to "rethink their teaching, rather than merely extend their existing practices" (Horn et al., 2017, p. 51). Many instructional coaching programs use evidence-based reflection 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) in curated video clips. Another program, 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, despite calls by researchers and university-based coaching programs, the use of evidence may be underutilized in teacher discussions about instruction due to competing logistical demands (Horn et al., 2017; Lefstein et al., 2020).

In fact, scholars have recently begun to describe the range of ways coaches and teachers work together (Gibbons & Cobb, 2017). Frameworks such as Boguslav's Systematic Framework for Describing Coaching Discourse (2024) identify the wide variety of discourse moves coaches use across different coaching activities. These moves can satisfy different purposes such as building trust (Bengo, 2016), modeling instruction (Saclarides & Munson, 2020), or setting goals (Kochmanski & Cobb, 2023). Some moves may be especially suited to sharing and eliciting reflection from evidence of observed instruction, such as making or eliciting observations and interpretations of specific classroom phenomena and interactions. Studying the moves coaches employ when utilizing evidence (and the moves they make when they don't) can help us learn more about the potential of automated feedback to support evidence-based reflection.

Instructional Artifacts in Coaching

Instructional artifacts such as observation notes, lesson plans, and classroom recordings are often used as sources of evidence to elicit reflection during teacher professional learning. 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 classroom recordings 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 recordings (or any artifact) does not automatically lead teachers to use evidence or engage in reflection. Instead, learning activities must be purposefully designed and facilitated to support teacher growth, for example by featuring activities that prompt reflection and link to teachers' learning goals (Brophy, 2003; Erickson, 2014).

Recordings can also be costly to use well. Although selecting recorded classroom events to analyze instruction can be a useful practice for coaches and teachers (Calandra et al., 2008), it can be very time-consuming and expensive. Capturing instruction requires equipment and setup, and turning a raw recording into a usable artifact involves reviewing, clipping, and aligning footage to coaching goals (Sherin, Linsenmeier & van Es, 2009; Gibbons & Cobb, 2017). This labor competes with coaches' and teachers' already limited time.

These demands, and the kind of evidence yielded, vary by the medium of recording. Video can be especially actionable when well curated, because it preserves visual context such as gesture, which carries meaning in mathematics instruction (Alibali & Nathan, 2012). But video is costlier to collect, carries a larger energy and storage footprint, raises greater privacy concerns, and can heighten teachers' self-consciousness about being recorded (Lepp et al., 2023). Audio, by contrast, is less intrusive and cheaper to collect. Although it omits visual context, audio foregrounds classroom discourse—a substantial and measurable dimension of instructional quality that is itself associated with student learning (Sedova et al., 2019; Authors, 2023a). These tradeoffs extend to automated analysis. Classroom discourse can already be transcribed and

analyzed from audio at scale, whereas automated interpretation of video's visual features, such as gesture, gaze and posture, remains comparatively nascent (Foster et al., 2024; Lee et al., 2026).

Regardless of the medium, however, the artifact alone does not determine its value for teacher learning. As noted above, recordings support reflection only when their use is purposefully designed to prompt interpretation and connect evidence to teachers' goals (Brophy, 2003; Erickson, 2014). Individual clips provide lots of rich information about particular moments but do not necessarily help teachers reason about connections among those moments (Coles, 2014). What matters more than the modality of capture, then, is how reflection on the resulting evidence is scaffolded. This is where coaching, and increasingly, automated tools that support coaching, can bring value by curating salient moments, surfacing patterns across lessons and structuring interpretive work (Matsumura et al., 2025).

Automated Feedback as Instructional Artifacts in Coaching

Researchers are increasingly turning to computational methods to make formative feedback on teaching more frequent and affordable (Jensen et al., 2020; Li et al., 2025). Advances in automatic speech recognition, natural language processing—including, most recently, large language models—now make it possible to analyze classroom recordings at scale, surfacing features of instruction such as teacher-student talk ratios, question types, and high-leverage talk moves or teaching and learning discourse practices like teachers' uptake of student ideas (Authors, 2022; Authors, 2024; Suresh et al., 2021). A growing set of tools, such as TeachFX (www.teachfx.com), TalkMoves (Suresh et al., 2021), and MPT (Demszky & Hill, 2023; Demszky & Liu, 2023; Demszky et al., 2025), analyze recordings of teachers' instruction across multiple subject areas. These tools highlight moments for reflection (e.g. talk moves) and

statistics (e.g. talk ratios) and using visuals such as timelines and charts, as illustrated in the next section. Initially, these feedback tools were envisioned for teachers' self-directed professional learning (Authors, 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 (Authors, 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 (Authors, 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. These findings suggest potential value in automated discourse-based feedback generated from audio recordings.

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 (Authors, 2025; Jacobs et al., 2022, 2024). Likewise, Jacobs et al. (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, scholars began to integrate these tools with instructional coaching.

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. Coaches can help build community-based structures and norms that position regular engagement with feedback as a shared, socially supported practice. In this integrated model, automated feedback becomes another instructional artifact that affords coaching conversations grounded in classroom evidence reflection.

Despite promising evidence that automated feedback can function as an instructional artifact for evidence-based coaching, important questions remain about feasibility, interpretation, and implementation. 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 MPT: asking what opportunities and challenges they anticipate, observing how they incorporate it with coaching moves in conversations with teachers, and documenting their reflections after initial use.

The MPT 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 M-Powering Teachers, (https://mpoweringteachers.stanford.edu) 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. 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; Authors, 2023a; Authors, 2021). The moves tagged by MPT serve as proxies for dialogic instruction, in which students are active participants of classroom discussions and teachers take up and build on their ideas (Nystrand et al., 1997; Alexander, 2018; Michaels, O'Connor & Resnick, 2008). The specific features MPT captures have themselves been linked to student outcomes (Authors, 2021; Authors, 2023a; Authors, 2024).

These computations are all done using locally-run models hosted on university servers. No data is exchanged with third-party AI companies.

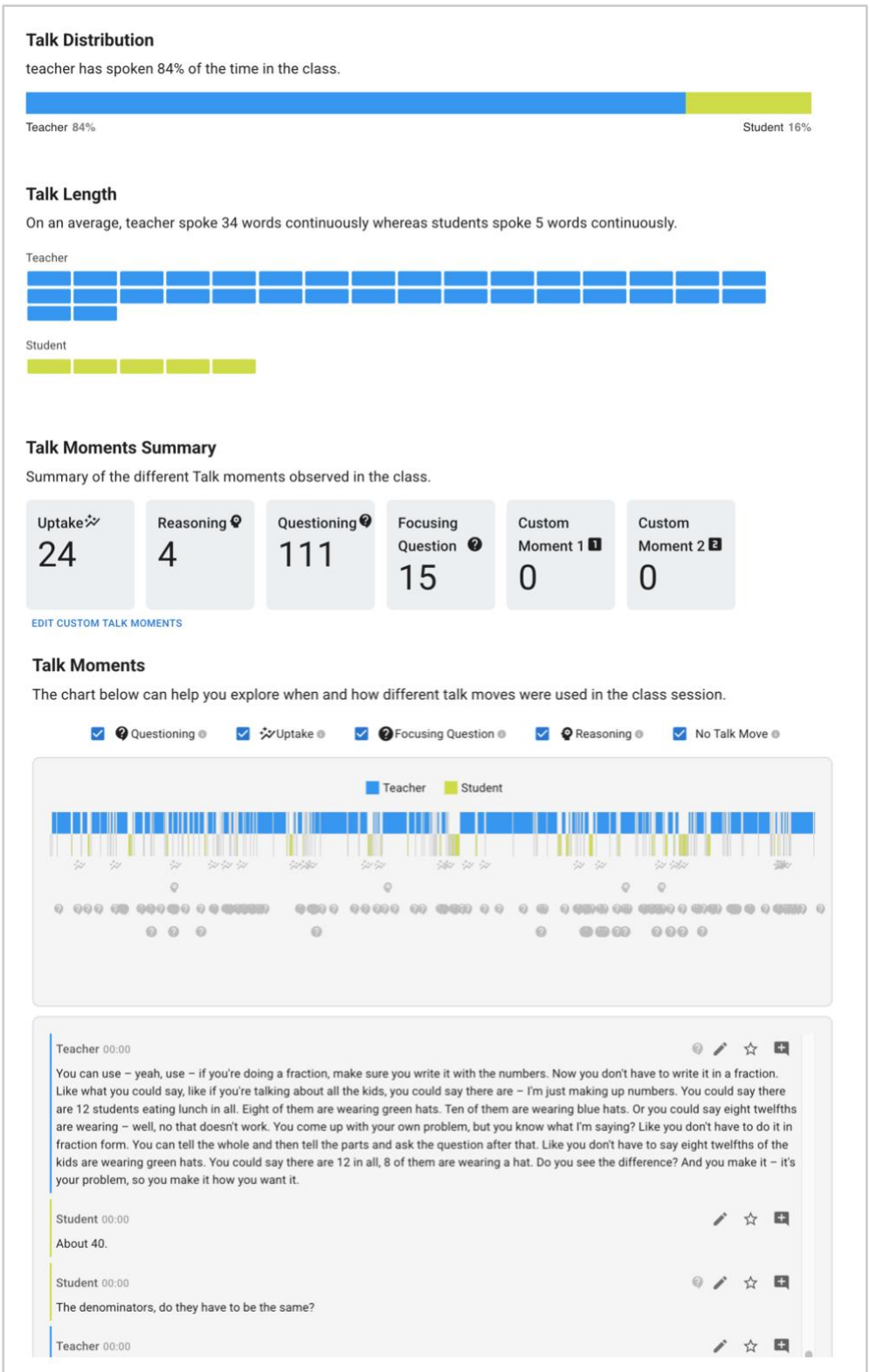
User Interface

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 continuous 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.

Before beginning this study, we interviewed 20 mathematics coaches and teacher educators to gather perspectives on a prototype of MPT. Suggestions from these interviews led to a few updates to the user interface of MPT to improve its utility and user-friendliness for instructional coaches (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).

Fig. 1 *Screenshot of the MPT Feedback Dashboard*



Methods

Analytic Sample

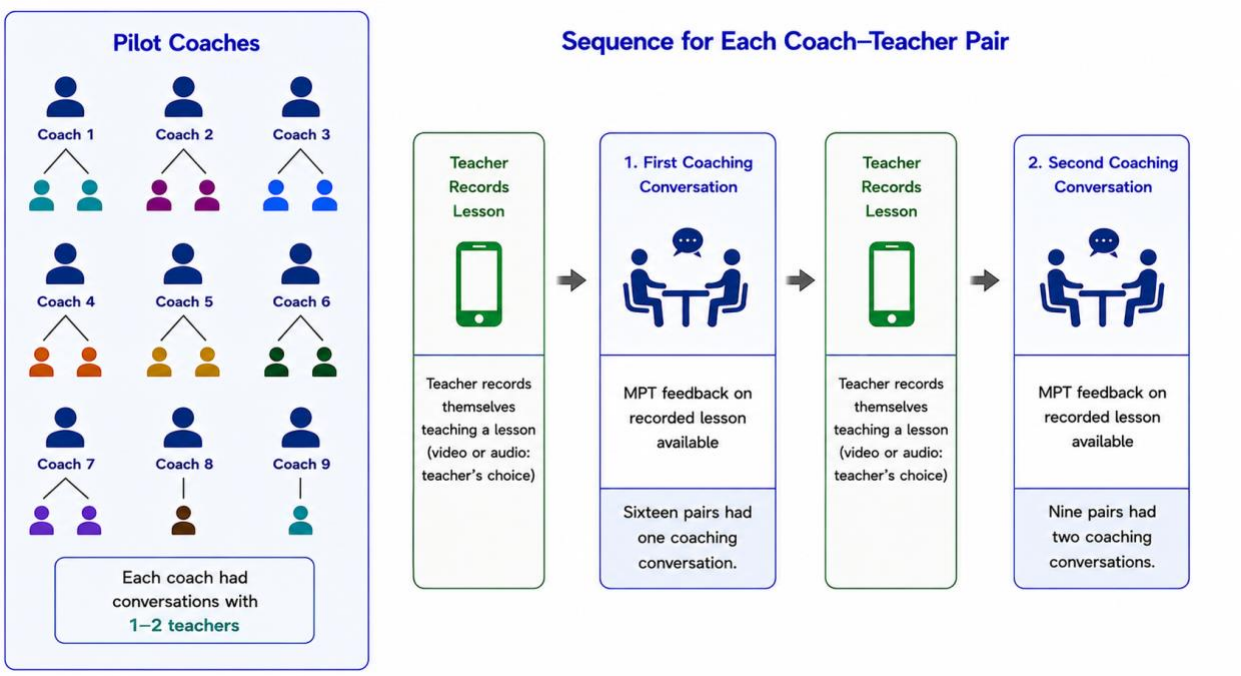
Our analytic sample includes 25 recorded coaching conversations of nine coaches working with 16 different teachers. Each coaching conversation lasted, on average, about 32 minutes. We also analyzed interviews with each of those nine coaches after they piloted MPT during coaching conversations.

Study Procedures

Piloting MPT During Coaching Cycles

We recruited 10 instructional coaches to allow us to observe their coaching sessions during the same school year (See Figure 2). These pilot coaches were selected to encourage diversity of perspectives toward using automated feedback technology (e.g., a range from enthusiasm to skepticism about MPT), 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 other coaches supported mathematics teachers or elementary teachers with the exception of one who coached a science teacher. 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). Coaches selected one to two teachers they were currently coaching or had previously coached, based on availability and willingness to try MPT, to participate in the pilot.

Fig. 2 *Overview of the Pilot*



As part of the pilot onboarding process, coaches 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 CSET and MQI to provide options and examples for how to use MPT in coaching conversations with teachers. These materials include suggestions for structuring coaching conversations such as specific reflection questions for each MPT metric and are available on Open Science Framework (https://osf.io/n5hr9/?view_only=bf24cfcac12b49648d8b7a5c55825fa6).

Researchers created user accounts on MPT for teachers and coaches. To receive MPT feedback, the teacher recorded a lesson they taught using a unique Zoom link provided by the researchers between coaching conversations. This recording was often made using a teacher's phone, iPad, or chromebook. To account for variation in recording quality across different

classrooms and avoid limitations of automated AI transcription technology at the time of the study, researchers sent the teacher-recorded audio to a human transcription service (transcribe.me) and then uploaded the transcript to MPT, where the resulting feedback could be seen by teachers. The feedback was available to them within 1 week (given the turnaround time of manual transcription), and teachers could then share the feedback with their coaches. Coaching sessions between coach and teacher pairs were recorded using Zoom, via a researcher-provided link.

Post-pilot Interviews

After the pilot, we invited coaches to participate in semi-structured reflective interviews. The interviews centered on coaches' experiences using MPT, the benefits they perceived, and their concerns. Coaches were also able to give feedback on the coaching training materials and routines. 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). We emailed each coach with a request to review an earlier draft of our findings. For those seven coaches that confirmed, we sent them a draft of the findings section along with a list of any examples or quotes that came from their observations or interviews. We asked if they had any general feedback, any specific feedback on their personal examples, or if they had any other questions or comments. They all confirmed that the framing was consistent with their perceptions and none offered any suggestions or revisions.

Analysis

To answer our first research question, we conducted an interaction analysis of coaching

conversations to better understand which coaching moves co-occurred with the use of automated feedback. To do so, we applied Boguslav's (2024) Systematic Framework for Describing Coaching Discourse in a comprehensive coding of every coach utterance across all recorded coaching conversations (see Appendix A). An utterance is defined here as a continuous stretch of speech produced by a single speaker, bounded by a point at which the floor may pass to another speaker (Sacks et al., 1974). Boguslav's framework, summarized in Table 1, is organized around two dimensions: temporal orientation (backward-facing, reflecting on past practice; or forward-facing, preparing for future instruction) and discourse type (asking questions versus making statements or directives). Note that some coaching moves, such as Justification, appear as both questioning ("Ask") and declarative ("Tell") forms.

Table 1 *Summary of Boguslav's Systematic Framework for Describing Coaching Discourse (2024)*

	Ask	Tell
Back	Interpretation, Self-assessment, Noticing, Cause & Effect, Vision, Justification, Lessons Learned, Grading	Interpretation, Positive Evaluation, Observation, Check In, Justification, Connection, Negative Evaluation, Grading
Forward	Generation, Goal Setting, Anticipation, Application, Check for Understanding, Content Understanding	Implementation, Instructional Strategy, Challenge, Demonstration, Student Goal, Reinforcement, Content Understanding

The same coach utterance could be coded with multiple moves. Each utterance was also coded for whether or not a coach directly utilized MPT during that utterance. We defined usage of MPT as any reference to specific MPT metrics, diagrams, or transcript excerpts while enacting a coaching move. Three members of the research team, each with more than 10 years of experience as instructional coaches and teacher educators, independently coded each coach

utterance across all recorded coaching conversations and discussed discrepancies until consensus was reached (Cascio et al., 2019). This coding allowed us to see the distribution of coaching moves used across 1,316 coach utterances by different coaches and in different coach-teacher pairs and which of those moves co-occurred with reference to MPT.

Some coaches incorporated MPT into the coaching activities they were already using, others dedicated a portion of their meeting time specifically for MPT, and others did both. In all cases, coaches varied from utterance to utterance in which specific coaching moves were utilized and whether or not they co-occurred with reference to MPT. We compare utterances employing the same coaching move but differing in whether MPT was used during the move.

We estimated a logistic regression model to examine associations between each coaching move and the likelihood that that move co-occurred with the utilization of MPT at the utterance level in coaching conversations:

$$Y_i = MPT_i \in \{0, 1\}$$

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{i27})$$

$$p_i = \Pr(Y_i = 1 \mid X_i)$$

$$p_i = 1 / [1 + exp(-(\beta_0 + \sum_{k=1}^{K} \beta_k X_{ik}))]$$

The binary outcome Y_i indicated whether there was reference to specific MPT metrics, diagrams, or transcripts in a coach's utterance i. Let p_i denote the probability that MPT was referenced in coach utterance i. The model specifies the log-odds of MPT reference as a linear function of a set of coaching moves where X_{ik} represents the presence of the kth coaching move for utterance i, and k denotes the 27 coaching moves coded in the analytic sample. The intercept β_0 represents the baseline log-odds of MPT reference when no coaching moves are present, and each coefficient β_k represents the change in the log-odds of MPT reference associated with the

presence of the corresponding coaching move, holding all other moves constant. That is, the models estimate how many times greater or smaller the odds of a coach utterance directly referencing MPT are when a specific coaching move is present (controlling for all other coaching moves).

Standard errors were clustered at the coach-teacher pair level to account for how individual coaches may favor some moves over others generally, when working with particular teachers, and when reflecting upon particular observed lessons. No problematic multicollinearity among coaching moves was observed (all VIF < 1.2). To address multiple hypothesis testing, p-values were adjusted using the Benjamini-Hochberg false discovery rate procedure.

The regression results suggested that MPT was often being used as a source of evidence for making observations and interpreting teachers' instruction (See Findings). Thus our second research question focuses on categorizing the *ways* in which coaches utilized automated feedback to share their own and elicit teachers' evidence-based reflection MPT. We conducted a second round of coding (Engle et al., 2014), where the same three members of the research team independently open-coded each utterance containing a coaching move that was positively associated with MPT use (see Findings) and wrote analytic memos about what they noticed and wondered. The research team then discussed these open codes and memos to reconcile them into four distinct themes (Strauss & Corbin, 1998).

To answer our final research question, we analyzed the post-pilot reflective interviews with coaches. These nine interviews were open coded by the research team (Given, 2008) into categories of affordances, concerns, and whether or not they thought "overall…that the automated feedback tool or coaching protocol 'worked' with the teachers [they] coached" and why. The research team discussed coding until consensus was reached.

Findings

Coaching Move Co-occurrence With Automated Feedback Use

In order to answer how coaches utilized automated feedback in coaching conversations, we first report the frequency of each coaching move and its co-occurrence with MPT during individual utterances made during coaching conversations (Figure 3). We then triangulate these frequencies using logistic regression models that predict the likelihood of MPT being utilized with each individual coaching move.

Fig. 3 Frequency of Coaching Moves With and Without MPT Reference

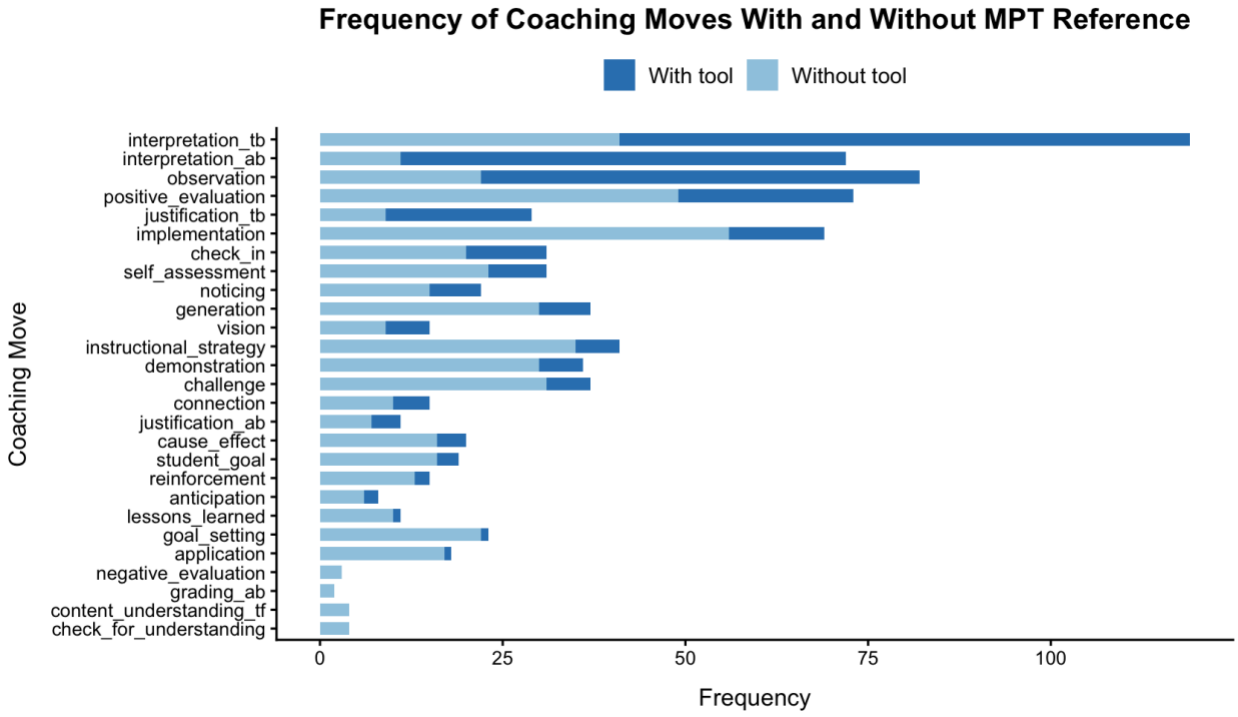


Figure 3 presents the frequency of each coaching move during coaching conversations. Because coaching moves were coded at the utterance level, an individual coach utterance may or may not specifically utilize MPT even when it was available during a given coaching conversation. The bars display different colors for when a move was used with or without reference to specific MPT metrics, diagrams, or transcript excerpts. The most frequently used moves overall were *Interpretation (Tell Back)* (n=119), *Observation* (n=82), *Interpret (Ask Back)* (n=72), and *Positive Evaluation* (n=73). These moves involved coaches making or asking the teacher to make an inference, describing specific factual information about an observed lesson, or communicating a positive judgement.

To understand coaching moves' co-occurrence with reference to MPT, we examined both frequencies and the significance of the association estimated by the logistic regression model. We found four moves, *Interpretation (Tell Back)*, *Interpretation (Ask Back), Observation,* and *Justification (Tell Back),* that frequently co-occurred with MPT references and whose association was also significant: *Interpretation (Tell Back)* (78 out of 119 moves, OR = 4.81, 95% CI [2.36, 9.80, $p < 0.001$], *Interpretation (Ask Back)* (61 out of 72 moves, OR = 13.57, 95% CI [3.91, 47.17], $p < 0.001$], *Observation* (60 out of 82 moves, OR = 6.20, 95% CI [2.70, 14.25], $p < 0.001$], and *Justification (Tell Back)* (20 out of 29 moves, OR = 5.16, 95% CI [1.58, 16.86], $p < 0.05$). As the numbers indicate, the majority of each move co-occurred during utterances which included MPT reference. Further, when holding other moves constant, they were at least 4 times as likely to co-occur with MPT reference than not.

Each of these moves involved coaches sharing or prompting teachers to share what they noticed or inferred from a prior lesson being discussed. Sometimes these moves were used during utterances that did not reference MPT, even during conversations in which it was

available. In utterances without MPT reference, these observations and interpretations sometimes drew upon coaches' prior knowledge of the teacher or students, their memories of or notes from the observed lesson, a teacher's narrative about what happened in the lesson, or video recordings. However, the overall pattern suggests that coaching moves emphasizing evidence-based reflection were most strongly associated with MPT reference.

We found one coaching move, *Challenge* (6 out of 37 moves, OR = 0.34, 95% CI [0.20, 0.58], $p < 0.001$), employed in utterances in which MPT was significantly less likely to be referenced. These moves contained "feedback that articulate[d] a challenge or problem of teaching." (Boguslav, 2024, Table 1, p. 14) and often manifested as coaches empathizing with some difficult aspects of teaching such as "You don't have a co-teacher or a para or anything" or "If [the students] haven't done it last year, they're not used to that, it takes practice." It is possible that these expressions of empathy drew more on prior experience than anything offered by automated feedback. However, there were some counter-examples in which MPT reference co-occured with utterances containing *Challenge* moves such as "Yeah, you could say [don't worry about the recording device] all day long but [students] are going to, perhaps, behave a little differently" or "So [this instance of uptake is] clarifying it for the other students which – you'd be amazed how much you need to do that in elementary and middle school but high schools still need [it]." Overall, however, coaches were less likely to reference MPT when using *Challenge*. This may indicate that the types of evidence provided by MPT is not conducive to this move. It is also possible that this move is better suited to supporting other non-evidence-based-reflection goals of coaching conversations (e.g. building trust).

Using MPT Evidence to Share and Elicit Reflection

The coaching moves *Interpretation (Askback)*, *Interpretation (Tellback)*, *Observation*, and *Justification (Tellback)* became the focus of our second round of coding. These moves occurred in 219 of 799 observed coaching utterances and each involved coaches sharing their own (i.e. telling) or prompting teachers to share reflections about the observed lesson. However, increased frequency of these moves does not necessarily imply higher quality. Fewer well-positioned moves might be more desirable than many superficial instances. To address the question of *how* coaches utilized MPT for evidence-based reflection while employing these moves, we identified the following four themes: (a) demonstrating effective practices; (b) probing; (c) challenging; and (d) comparing and generalizing. We discuss each below and elaborate on two other coded themes, the use of custom annotation and utilizing tool inaccuracy.

Coaches were observed using MPT to demonstrate effective practices to teachers, 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 draw 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. Probes could also direct 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 features into subcategories—such as subcoding focusing questions as “why,” “what,” or

"how"—to examine how different focusing 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 leveraged the tool's limitations as a springboard for reflection 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.

Coaches' Conclusions about Automated Feedback

After piloting MPT, all but one of the nine coaches affirmed benefits relating to the use of evidence in coaching conversations (see Table 3). 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."

Most pilot coaches expressed favorable views of what they perceived as the objectivity of the evidence 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.

Table 2 *Affordances and Concerns of 9 Pilot Coaches*

Affordance	n/9	Concern	n/9
Encouraging more or **better teacher reflection**	8	**Not as useful** to more **experienced** teachers	2
Visual representations of abstract patterns	8	Limit **types of goals** teachers can set	2
Objectivity and specificity of evidence	7	General **inaccuracy**, especially of student talk	8
Saving time or creating **efficiency** for coaches and/or teachers	5	Wasting time or **creating unproductive extra work** for coaches and/or teachers	2
Ability to **revisit specific moments** using transcript	3	Incentivizing **unintended consequences**	2
More teacher **access to coaching**	2	**Lack of integration** with schools' existing teacher observation protocols	1
Alignment with other professional learning initiatives	2	Feedback can be **overwhelming** for teachers	5
Entry point for teachers reluctant to be video recorded	1	**User interface** and protocols could be clearer	2

A concern expressed by almost every piloting coach was the inability to accurately capture evidence of student talk, especially in small groups. This was not because of any limitations of the tool itself, but rather the devices used to record lessons (e.g. a teacher's phone or iPad). 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 recording quality of inputs into 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 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.

We also asked each coach if they thought MPT "worked" with the teachers they coached (Table 4). Five coaches reported that it did, two said maybe, and two said no.

Table 3 *Coach responses to "Overall, do you think that the automated feedback tool or coaching protocol "worked" with the teachers you coached? Why or why not?"*

Coach	Yes	Maybe	No	Rationale
Coach 1	✓			Teachers were excited, and the process felt less evaluative than typical coaching.
Coach 2	✓			The tool helped teachers set specific goals and change classroom behavior aligned with those goals.

Coach 3	✓			The tool was particularly useful for teachers reluctant to be video recorded.
Coach 4	✓			Observed changes in teachers' practice that could be partially attributed to the tool.
Coach 5	✓			Effective with teachers who were highly open to feedback; may be less effective with more reluctant teachers.
Coach 6		✓		Helpful for supporting reflection with a novice teacher but less impactful for an experienced teacher.
Coach 7		✓		Potential risk of misuse; could be harmful if used evaluatively rather than reflectively.
Coach 8			✓	Recording quality needs improvement before the tool could be effective.
Coach 9			✓	Would not use the tool without the ability to capture more student talk.

The reasons coaches gave for and against MPT's efficacy corroborate findings from observations of coaching conversations and highlight broader challenges that must be addressed for automated feedback tools to effectively support evidence-based reflection. These challenges include the need to improve transcription accuracy, particularly in capturing student talk; to ensure that feedback remains formative and actionable rather than evaluative; to support the development of rapport between teachers and coaches, especially for teachers who may be reluctant to engage in coaching; and to consider differentiated use cases that account for varying levels of teacher experience and openness to feedback.

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 and the coaching moves most predictive of MPT use were related to recalling and interpreting factual information from observations. When deemed accurate by coaches, we observed coaches using MPT 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.

This trust-distrust dichotomy speaks to a broader conversation on teachers' trust in AI generated evidence (Nazaretsky et al., 2022). The "objectivity" coaches attributed to MPT may reflect tendencies towards automation bias, in which people treat AI outputs as more authoritative than warranted and may over-rely on them (Parasuraman & Riley, 1997). We also observed the opposite tendencies—dismissing MPT because of imperfect transcription, which

can lead users to discard genuinely useful information (Dietvorst et al., 2015). The most productive interactions we observed reflected something in between, what some might call "appropriate reliance" (Lee & See, 2004), in which coaches positioned MPT as "another pair of eyes," a fallible participant whose claims were to be examined, and modeled for teachers how to weigh automated evidence against their own professional knowledge. As automated tools become more prevalent in education, cultivating this kind of calibrated trust may be coaching's most valuable contribution.

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 over tested. 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 of what may be exceptionally skillful university program-affiliated coaches. 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. Additionally, the inputs to MPT for this study were made a human transcription service. Automated transcription of classroom recordings may amplify concerns listed above about transcription accuracy.

To better realize the potential of MPT and similar tools as instructional artifacts for evidence-based teacher reflection, 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). As these metrics are developed, more qualitative study will be needed into how coaches and teachers utilize them to cultivate evidence-based reflection.

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 (see Foster al., 2024), 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 utilized 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 reflection, 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 reflection in support of teacher learning and student understanding.

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alexander, R. (2018). Developing dialogic teaching: Genesis, process, trial. *Research Papers in Education, 33*(5), 561–598.

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences, 21*(2), 247–286.

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. Zesch (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://doi.org/10.18653/v1/2022.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/jresematheduc-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). *Using video in teacher education* (Advances in Research on Teaching, Vol. 10). 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 Research. (2023). *MQI coaching cycle*. Harvard University. https://mqicoaching.cepr.harvard.edu/mqi-coaching-cycle

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

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. Zesch (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). Association for Computing Machinery. https://doi.org/10.1145/3573051.3593379

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). 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). Measuring conversational uptake: A case study on student–teacher interactions. In C. Zong, 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* (Vol. 1, 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

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

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

Foster, J. K., Korban, M., Youngs, P., Watson, G. S., & Acton, S. T. (2024). Automatic classification of activities in classroom videos. *Computers and Education: Artificial Intelligence, 6*, Article 100207. https://doi.org/10.1016/j.caeai.2024.100207

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. https://doi.org/10.1007/s11858-010-0280-7

Gibbons, L. K., & Cobb, P. (2017). Focusing on teacher learning opportunities to identify potentially productive coaching activities. *Journal of Teacher Education, 68*(4), 411–425.

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

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

Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery.

Kochmanski, N., & Cobb, P. (2023a). Identifying and negotiating productive instructional improvement goals in one-on-one mathematics coaching. *Journal of Teacher Education, 74*(5), 437–450.

Kochmanski, N., & Cobb, P. (2023b). Identifying productive one-on-one coaching practices. *Teaching and Teacher Education, 131*, Article 104188. https://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

Lee, H. H., Sung, H., Celedón-Pattichis, S., & Pattichis, M. S. (2026). Toward an inclusive understanding of collaborative learning using MMLA: Exploring multimodal participation dynamics by gender and linguistic diversity. *Learning and Instruction, 102*, Article 102317. https://doi.org/10.1016/j.learninstruc.2026.102317

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

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

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

Lepp, L., Leijen, A., Küüsvek, A., & Kalk, K. (2023). Teacher education students' first-time experiences of video-recording their teaching and analyzing it. *SAGE Open, 13*(3), Article 21582440231196746.

Li, K., Wang, P., & Chen, G. (2025). How can AI be integrated into teacher professional development programs? A systematic review based on an adapted technology-based learning model. *Teaching and Teacher Education, 168*, Article 105219.

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

Matsumura, L. C., Walsh, M., Li, T., Correnti, R., & Litman, D. (2025). Exploring the use of generative AI to identify "coachable moments" in classroom discussions. In *Proceedings of the 19th International Conference of the Learning Sciences–ICLS 2025* (pp. 2971–2972). International Society of the Learning Sciences.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283–297.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology, 53*(4), 914–931.

Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom*. Teachers College Press.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253.

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

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.

Saclarides, E. S., & Munson, J. (2021). Exploring the foci and depth of coach-teacher interactions during modeled lessons. *Teaching and Teacher Education, 105*, Article 103418.

Sedova, K., Sedlacek, M., Svaricek, R., Majcik, M., Navratilova, J., Drexlerova, A., & Salamounova, Z. (2019). Do those who talk more learn more? The relationship between student classroom talk and student achievement. *Learning and Instruction, 63*, Article 101217.

Sherin, M. G., Linsenmeier, K. A., & van Es, E. A. (2009). Selecting video clips to promote mathematics teachers' discussion of student thinking. *Journal of Teacher Education, 60*(3), 213–230.

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/

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006.

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*, 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: Frequency of Utterances Containing Coaching Moves (see Boguslav, 2024) with MPT Co-Occurrence

Coaching Move	Example	Freq. with MPT reference	Total Freq.[a]	Category Totals
Interpretation (AskBack)	"Do you feel like [the students] were having good conversations?	61	72	**Ask Back** MPT n = 91 Total n = 184
Self-assessment	"What else can we celebrate that you saw happening here?"	8	31	
Noticing	"What did you notice about your application problem today?"	7	22	
Cause & Effect	"How [have the sentence starters] been working for that class?"	4	20	
Vision	"Before we start, I wanna first look at what is your focus?"	6	15	
Justification (AskBack)	"What made you check in individually with that one student?	4	11	
Lessons Learned	"Any key learning or takeaways from the experience this year?"	1	11	
Grading (AskBack)	"What were some reinforcement areas or things you did well [according to this rubric]?"	0	2	
Generation	"What's something we can do to help you with delivering that part of the lesson?"	7	37	**Ask Forward** MPT n = 11 Total n = 90
Goal Setting	"Let's talk about what you wanna do next in terms of focusing questions"	1	23	
Application	"The quick check isn't for the math centers, [it's] just for the day?"	1	18	
Anticipation	"Do you think they'll understand the white feathers vs. the black feathers?"	2	8	
Check for Understanding	"I wanna hear from you what comes to mind after reading what reasoning looks like and what we should be looking for."	0	4	
Content Understanding (AskForward)	-	0	0	
Interpretation (TellBack)	"There was nothing overly complex in terms of your instruction to them"	78	119	**Tell Back** MPT n = 198 Total n = 352
Observation	"You asked a focusing question, a student gave a response, and then you talked"	60	82	
Positive Evaluation	"It looks like every instance of student reasoning had an uptake. So that's awesome!"	24	73	
Check In	"Last time we had a plan…to work on student reasoning"	11	31	
Justification (TellBack)	"When you asked these questions, you wanted students to explain to you what they did"	20	29	

Connection	"I think [the student response] had a lot to do with your framing"	5	15	
Negative Evaluation	"I didn't see a lot of examples in here where you asked other students to revoice"	0	3	
Grading (TellBack)	-	0	0	
Implementation	"If they had all the problems…they're all gonna have a period of time to consider the problem independently"	13	69	**Tell Forward** MPT n = 36 Total n = 221
Instructional Strategy	"Once the timer goes off, that's the opportunity where you can bring in that student ownership"	6	41	
Challenge	"It takes practice…if they haven't gotten it last year they don't just immediately [talk that way]"	6	37	
Demonstration	"[Students will] say 'Oh because my 10 is different than his 10.' [Then you could say] 'But you both had 10s. Why is that not the same?"	6	36	
Student Goal	"So your next action step…will be making sure you have the cards for their boards."	3	19	
Reinforcement	"So continue pushing for those sentence starters"	2	15	
Content Understanding (TellForward)	"That is super helpful in future math because when we're doing rational equations you have expressions like…"	0	4	