International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00417-x

ARTICLE

High School English Teachers Reflect on Their Talk: A Study of Response to Automated Feedback with the Teacher Talk Tool

Sean Kelly[1] · **Gizem Guner**[1] · **Nicholas Hunkins**[2] · **Sidney K. D'Mello**[2]

Accepted: 25 June 2024

Abstract

We present the Teacher Talk Tool, which automatically analyzes classroom audio and provides formative feedback on key aspects of teachers' classroom discourse (e.g., use of open-ended questions). The tool was designed to promote teacher learning by focusing attention and sense-making on their discourse. We conducted a feedback-response study where five English & Language Art teachers used the Teacher Talk Tool in eight classroom sessions. Teachers completed repeated-measure surveys and semi-structured interviews providing quantitative and qualitative evidence of feedback response. Results indicated that the majority of automated feedback was perceived to be accurate and prompted a high degree of reflection, focusing teachers' attention on the measured talk constructs. This feedback also led teachers to engage in a process of sense-making, linking the measured talk features to classroom processes and contexts. However, evidence of feedback uptake was more limited. Overall, results contribute to the nascent literature on the efficacy of automated feedback on instructional practice.

Keywords Instructional observation · Teacher learning · Classroom discourse · Automated feedback

✉ Sean Kelly
spkelly@pitt.edu

Gizem Guner
gig23@pitt.edu

Nicholas Hunkins
Nicholas.Hunkins@colorado.edu

Sidney K. D'Mello
sidney.dmello@gmail.com

[1] Department of Educational Foundations, Organizations, and Policy, University of Pittsburgh, Pittsburgh, PA, USA

[2] Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA

Published online: 08 July 2024

Introduction

A good teacher is first and foremost a good learner. Due to the contextualized nature of teaching, a cycle of planning, enacting, reflecting, and adjusting lies at the core of becoming a good teacher (Clarke & Hollingsworth, 2002). Learning also requires feedback (D'Mello et al., 2010; Azevedo & Bernard, 1995; Ericsson et al., 1993; Shute, 2008). Yet, teachers deliver approximately 900 "performances" (or class sessions) a year (for a secondary school teacher with five classes and 180 days of school), with few opportunities to pause and receive feedback from a trusted source (Fadde & Klein, 2010; Palonsky, 1986; Stigler & Miller, 2018). This lack of immediate and objective feedback is a critical barrier that must be overcome to enhance teacher learning.

We introduce the Teacher Talk Tool (3 T), an artificial intelligence in education (AIEd) system providing automated, non-evaluative feedback on classroom discourse, an important dimension of instruction affecting student engagement and learning (Caughlan et al., 2013; Gamoran et al., 1995; Kelly, 2007; Kelly & Abruzzo, 2021; Langer, 2001; Murphy et al., 2009; Reznitskaya et al., 2001; Taylor et al., 2005). Artificial Intelligence in Education (AIED) systems include tools designed for a wide variety of instructional contexts and activities (see e.g., Datta et al., 2023; D'anjou et al., 2019; Gerard et al., 2020; Sankaranarayanan et al., 2020; Tran et al., 2023). The Teacher Talk Tool is part of the subset of AI-based technologies that combine sensing technology (e.g., video-cameras, microphones, wearables, etc.) with computational methods including machine learning to measure and classify aspects of teaching and learning during interactive classroom instruction (see e.g., Ahuja et al., 2019; Demszky, 2022; Huang et al., 2020; Jacobs et al., 2022). In this case, we aim to encourage teachers to reflect, in particular, on their own discourse. Such data-driven reflection is an effective strategy for improving teacher effectiveness via job-embedded professional development efforts (Camburn, 2010; Camburn & Han, 2015; Putnam & Borko, 2000). Automation offers the potential for radically more efficient and self-directed observation and feedback. Gains in efficiency accompanying automation are particularly noteworthy when the goal is fine-grained observation of instruction (e.g., categorization of each utterance, seconds of time use, etc.). Yet, automation also fundamentally changes the ecology of feedback, affecting teachers' response to feedback in unknown ways (see Section "The Ecology of Traditional Feedback").

The current tool builds on our prior research demonstrating the reliability and validity of automated observation of classroom discourse in English and Language Art classes (Jensen et al., 2020; Kelly et al., 2018). For the first time, in this study we provided teachers with state-of-the-art, fully-automated feedback from classroom audio data, including several constructs not found elsewhere in the nascent literature on automated instructional feedback (Demszky, 2022; Jacobs et al., 2022). The discourse constructs in 3 T extend Gamoran and Nystrand's (1992) program of research, with additions inspired by the Protocol for English Language Arts Teaching Observation (Grossman et al., 2013) and Shernoff's (2013) model of Environmental Complexity. However, specific constructs found

in our prior work to be difficult to reliably automate with existing technology or rarely occurred (e.g., uptake) were not included.[1] Although our focal dimensions of effective discourse (see Table 1) are drawn from these three frameworks, we recognize that similar concepts are central to over-arching models of discourse, such as Accountable Talk (Resnick et al., 2018), Quality Talk (Wilkinson et al., 2010), and Questioning the Author (McKeown & Beck, 2015).

The feedback we provided was not designed to provide a summary evaluation; talk constructs are not defined by effectiveness, and no overall judgement of effectiveness is made. Further, our feedback is based on especially well-validated automated procedures (Jensen et al., 2021). Prior research has validated feedback algorithms only on human-transcribed audio (Lugini et al., 2019; Suresh et al., 2018) or a combination of computer + human transcriptions (Song et al., 2021), which impedes scalability and timeliness of feedback. In other cases, research occurred in online contexts where speech diarization (segmentation/identification of who spoke when) is especially accurate (Alic et al., 2022). Further, in some cases, prior testing procedures do not support claims of generalizability (e.g., Song et al., 2021; Suresh et al., 2018) because they do not ensure teacher-level independence in training and testing sets.

In both traditional and automated forms of feedback, teacher's acceptance of and willingness to use feedback should not be assumed (Quintelier et al., 2020). Thus, our overall study aim is to investigate a set of conceptually-related responses to feedback that begin to speak to the efficacy of automated feedback on instructional practice. First, do teachers find the automated feedback accurate and important? (RQ1) Second, does the provision of automated feedback improve teacher's attentiveness to classroom discourse, focusing attention and generating related sense-making? (RQ2) Third, do teachers show evidence of the uptake of feedback, including goal setting and strategy use? (RQ3) Collectively, these outcomes speak to a set of conceptually interrelated and consequentially valid feedback responses (Brett & Atwater, 2001; Chawla et al., 2019; Quintelier et al., 2020) discussed further in Section "Summary Conceptual Model & Research Questions".

To address these questions, we conducted a feedback-response study where five teachers used the tool to receive automated feedback on both teacher-led discourse and transactional (i.e., student–teacher) discourse in eight lessons each. Teachers completed self-report items pertaining to the feedback and were interviewed twice. The study adopted a minimal-intervention strategy of providing teachers with a passive AIED technology for self-directed use and reflection without any external scaffolding (see discussion in Chiu et al., 2022), thereby providing a strong test of feedback response among the limited number of study participants. For contrast, as an example of an *intensive* technology-enabled (but not fully automated) intervention, Chen et al.'s (2020) video-based professional development workshop study provides

[1] The robustness with which uptake is identified may depend strongly on the specific definition and context of uptake. In our prior work in English language arts contexts, we employ Nystrand and Gamoran's very strict definition of uptake. Other researchers have had success coding more expansive treatments of uptake in computer science classrooms (Demszky et al., 2023).

Table 1 Discourse features in the Teacher Talk Tool

	Definition	Example
Instructional Talk	Talk focusing on the lesson and learning goals rather than on other topics, such as classroom management, procedural talk (e.g., "get out your pencils"), or other talk not related to the lesson	"Let's discuss the theme of the novel"
Open-Ended Questions	Open-ended questions are those whose answers are not pre-specified by the teacher. They position students as having knowledge that the teacher does not, so some educators refer to them as "authentic questions."	"Why do you think this is the most important part of the story?"
Elaborated Feedback	Elaborated feedback expands on student comments by providing reasons why the comment was strong, interesting, flawed, etc. and provides explicit guidance for student learning and thinking	"That's true, our character may be an anti-hero because he seems burdened by helping others but he continues to help."
ELA Terms	The disciplinary terminology of English language arts, which helps students to use the language and concepts of ELA in their learning. For example, students "see" more foreshadowing when they have a term for this literary device	"In your essays, each paragraph should contain a claim, evidence to support your claim, and reasoning or why your evidence proves the claim."
Goal Clarity	Explanations of learning goals and procedures for activities that help students understand why and how they should participate in a particular activity or task	"Your writing partner should give you three overall comments, before any minor editing stuff."

strong evidence of instructional change. The present study was intended to address basic questions pertaining to teachers' usage and perceptions of the tool, paving the way for future, larger studies on efficacy.

The Ecology of Traditional Feedback

Much more is known about how teachers perceive and respond to traditional in-person observation structured by global protocols (Kelly et al., 2020a, b) than about how teachers react to automated feedback. Global observation protocols, like those found in the Framework for Teaching (FFT), the Marzano Focused Teacher Evaluation Model, and the TAP System for Teacher and Student Achievement provide rough, qualitative, but richly comprehensive assessments on numerous core dimensions of teaching. The goal of these protocols is to structure judgements of multiple aspects of teaching (e.g., the four sub-domains of Domain 2 of FFT are: Creating an Environment of Respect and Rapport; Establishing a Culture for Learning; Managing Classroom Procedures; and Managing Student Behavior), which combine to form an overall rating of teaching effectiveness. Such observation systems are in widespread use around the United States for teacher evaluation and professional development (Close et al., 2018; Wieczorek et al., 2022), and have been developed and adapted for use in many other countries (Klette et al., 2017; van de Grift 2014; White & Klette, 2023).

Considering in-person observation with global protocols compared to automated, fine-grained systems like the one investigated in this study: global protocols offer a more comprehensive assessment of instructional practice (Praetorius & Charalambous, 2018), but only rough, qualitative distinctions instead of fine-grained measurement (Hennessy et al., 2020), and are more costly to carry out (Archer et al., 2016). These systems are also, by definition, highly evaluative. Beyond these basic, surface level properties, understanding of the measurement properties of global observation protocols has been supported by major studies including the circa 2010–2012 Measures of Effective Teaching Study. This literature suggests numerous difficulties that limit the robustness of teacher evaluation and feedback, limiting the overall quality of the information gleaned from observation (Bell et al., 2014; Campbell & Ronfeldt, 2018; Cohen & Goldhaber, 2016; Gitomer et al., 2014; Humphry & Heldsinger, 2014; Kelly et al., 2020a, b; Liu et al., 2019; McCaffrey et al., 2015; White, 2018). These challenges include: a lack of independence in sub-domains of instruction that are needed to provide teachers useful, domain-specific feedback (Aucejo et al., 2022; Humphry & Heldsinger, 2014; Liu et al., 2019; McCaffrey et al., 2015); dramatic differences in reliability depending on rater training (Kelly et al., 2020a, b); a tendency for the vast majority of ratings to cluster in the middle of the rating scale (Kelly et al., 2020a, b; Kraft & Gilmour, 2016). These limitations, if known or sensed by teachers, could affect teachers' overall sentiments about feedback. On the other hand, while there is substantial variation in the inclusion and labeling of constructs across popular protocols (Praetorius & Charalambous, 2018), we suspect many users view the global protocols selected for use in their districts as generally coherent and well-developed instantiations of best practices in their

disciplines. Overall though, with such little information to go on besides face validity, we wonder how positively teachers evaluate the tools that are used to scrutinize their instruction?

In a major study of teacher response to observational evaluation and feedback, Cherasaro et al. (2016) report very mixed perception of overall usefulness. Seventy percent of teachers viewed their evaluator (most often principal or assistant principal) as credible, and 74% viewed the feedback as accurate, but only 55% agreed or strongly agreed the feedback was useful. Yet, from this study, we do not know how features of the observational protocols themselves affected teacher positivity, or how the typically highly kurtotic (as opposed to more uniform) score distribution affected positivity. Moreover, the Cherasaro et al. (2016) study and others on this topic (Dwyer & Stuffelbeam, 1996; Grissom et al., 2018; Kraft & Christian, 2019; Pepper et al., 2015) occurred in the context of evaluation. It's possible teachers might be much more positive in a less evaluative context. On the other hand, even without an explicit evaluative use or context, teachers may be skeptical that an outside observer can appreciate the situated nature of their instruction or the need to adapt to their students. Certainly, the quality of the teachers' relationship with the specific observer, who is often a supervisor, is important (Quintelier et al., 2020). Studies of teacher-principal trust suggest high variability, and that many teachers are likely vulnerable to feelings of mistrust toward the administrators they work with (Price, 2012, 2021; Tschannen-Moran & Hoy, 1998; Van Maele & Van Houtte, 2009).[2] Overall, there is the concern that systems of teacher observation potentially face a double threat: mistrust of the accuracy and relevance of the tool or observational system, combined with a precarious state of teacher professional autonomy to begin with.

Emerging Research on Automated Feedback

Automated systems replace the credibility of and trust in individual, specific evaluators with the credibility of the automation and the features of the tool itself. Logg et al. (2019) argue that in the modern era, individuals have a surprisingly high appreciation for algorithmic predictions and estimations. In fact, their specific experimental evidence gathered in hypothetical scenarios shows greater adherence to algorithmic advice than that of a person. Perhaps most relevant to the current topic, Logg et al. (2019) even find respondents preferred algorithmic estimates to their *own* estimates (Experiment 3). There is a literature specifically on over-trust in automation (Aroyo et al., 2021), and a much larger literature on trust more generally, including skepticism (Hoff & Bashir, 2015; Schaefer et al., 2016). Yet, this research is not specific to teachers' professional judgement, and differs substantially in a particular way from the present topic of automated instructional observation. Automated teacher observation relies not just on predictive algorithms, but also on sensory input (Ahuja

[2] The studies cited here are focused more on the determinants of trust than characterizing the central tendency in trust, which would be aided by a frame of reference (i.e., common measures used with respondents in multiple occupations) not present in the data.

et al., 2019), a microphone accurately recording and transcribing teacher speech in the case of the current automated method. Thus, there are two major steps at which the teacher observation process could fail, whereas the human and algorithmic estimates shared identical inputs in the Logg et al. experiments.

Jacobs et al. (2022) and Demszky (2022) provide important evidence on how teachers respond to automated feedback, including estimates of the ratio of student to teacher talk in their classrooms among other constructs. The systems used in these studies, Talk Moves and the TeachFX system respectively, share many basic features with the present Teacher Talk Tool (described below in detail). Both systems are based on fine-grained measurement of individual utterances and other very precise acoustic and linguistic units of measurement. Both systems are agnostic insofar as judgements of effectiveness, appropriateness, etc., of a given talk move are not made at the time of coding (see Kelly, 2023 for further discussion). Additionally, neither system features an overall score, and certainly not an overall judgement of effectiveness—only scores for individual features are reported.[3]

There is some evidence in the Jacobs et al. (2022) study of the challenging ecology of automated feedback. Perceptions of accuracy of the TalkMoves feedback were highly variable, with about 20% of the teachers deeming the feedback inaccurate and the remainder having at least some reservations about accuracy. In this case, it is difficult to know whether users were overly skeptical (or perhaps just reasonably skeptical), because it is unclear if/how the scoring accuracy was communicated to the teachers. Further, Jacobs et al. (2022) did not report accuracies of the underlying scoring algorithms based on automatic speech recognition (ASR) technology inputs (i.e., the use case), reporting instead results on high-quality human transcripts (Suresh et al., 2022; Suresh, et al., 2019). ASR technology is notoriously inaccurate under noisy, real-world conditions (Blanchard et al., 2015; Cao et al., 2023; D'Mello et al., 2015; Southwell et al., 2022).

Additionally, teachers focused much more on one talk feature, the ratio of student to teacher talk than the others (this finding discussed more later). One reason for difficulty in perceptions of accuracy in this study may have been that the system goals were so ambitious, analyzing both student and teacher talk with a complex array of microphones. Yet, even with respondents having some reservations about accuracy, there was still much evidence of positive use of the system and teacher learning (see changes described on p. 8). Likewise, Demszky (2022) evaluated one component of TeachFX's feedback intervention, finding that instructors in an online undergraduate science course who received practice-based feedback on the incorporation of student ideas into instruction improved their use of this high-leverage teaching practice and improved student learning outcomes. In this prior work, we would posit that the agnostic (i.e., judgment-free) nature of the feedback strongly helped promote use.

[3] In the case of the Teacher Talk Tool the feedback did entail a comparison (to normative data), but because the talk constructs are not defined with reference to effectiveness, participants were not forced to infer any given comparative score was "good" or "bad." There are also some basic system differences to Jacobs et al. (2022): the features themselves are different, the audio recording systems are very different, and Jacobs et al. (2022) provided feedback on a continuous (0–100%) scale. Differences in the underlying computational models and their validation are not discussed here.

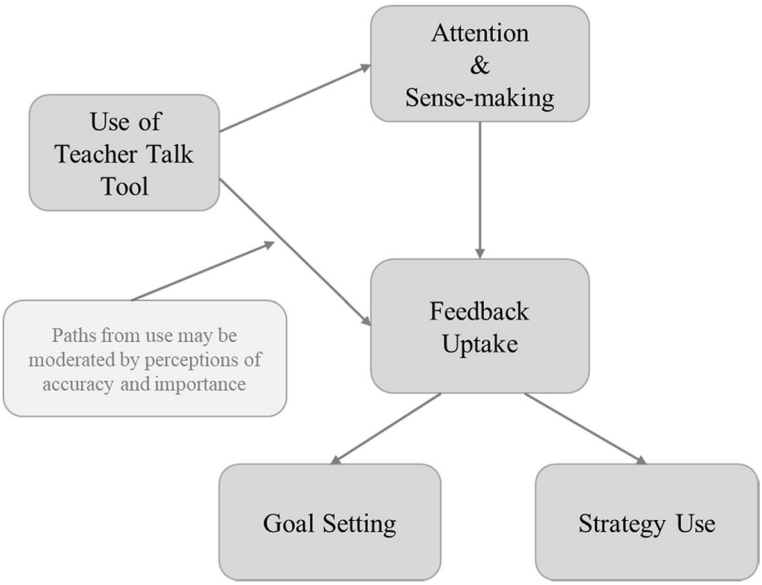

Fig. 1 Conceptual model of teacher discourse feedback response. Note: Differences in teacher emphasis on individual constructs not shown here

When judgements of effectiveness are made at the moment of observation, a natural reaction is that the observer or system does not understand the contextualized nature of the class, "my students' needs," "my school context," "my goals," etc. When judgement is withheld such reactions are not needed.

Summary Conceptual Model & Research Questions

Based on prior studies showing how feedback features affect user response (Brett & Atwater, 2001; Chawla et al., 2019; Quintelier et al., 2020), the goal of this study was to explore three inter-related responses to the Teacher Talk Tool: (RQ1) Perceptions of accuracy and importance; (RQ2) teachers' attentiveness to classroom discourse (an attention-focusing effect and related sense-making) and (RQ3) uptake of feedback including goal-setting and strategy use. Note that these are temporally distinct, such that users first carry out a classroom recording (which unto itself might focus teachers' attention on discourse), and then receive feedback (where the attention-focusing and sense-making might mediate uptake or behavior change in response to the feedback). Sense-making is a term from the teacher noticing and learning literature (see e.g., Colestock & Sherin, 2009) and refers here to teachers' personalized foci and interpretations of feedback based on their experience and perspective—how teacher's "read" feedback. Additionally, perceptions of accuracy and importance are not only basic dimensions of feedback response, but might moderate the relationship between tool use and more distal learning outcomes. These questions are captured in our feedback-response study conceptual model (see Fig. 1).

Following findings by Jacobs et al. (2022) we expected that teachers might focus more on some talk constructs than others; thus we investigated differences in teacher interest across the five focal talk constructs assessed by the tool.

Our paper makes two new contributions. First, we present (for the first time) the Teacher Talk Tool as an exemplary implementation of an AI-driven automated feedback system on the quality of teacher discourse in authentic classroom environments, a culmination of more than three decades of research on this topic. Second, we conducted a longitudinal feedback response study to investigate teachers' perceptions and use of the tool with respect to our conceptual model.

Data and Methods

Overall Study Design

The overall study design aimed to produce rich information on feedback response from teacher participants working in their real-world classrooms. To that end, we observed five teachers for eight lessons each (N = 40 lessons), with short, online surveys following each lesson recording, as well as each time the teacher received lesson feedback. Additionally, teachers participated in two interviews (one after Lesson Three and one after Lesson Eight), and completed pre- and post-study teacher questionnaires. Analytic methods include a mixture of quantification of repeated-measure survey data, coding and quantification of accuracy responses in the face-to-face interview data and further qualitative appraisal of the interview data.

Teacher Talk Tool

The Teacher talk tool provides independent teacher users with reliable feedback on five discourse practice variables (features): instructional talk, open-ended questions, elaborated feedback, disciplinary terminology in English and Language Arts, and goal clarity. These features are common to instructional observation systems, but are not exhaustive of all instructionally-relevant dimensions of discourse; further constructs could be added if measurement studies confirm they can be robustly identified automatically. The system design is currently tailored for use in English classrooms and focuses only on teacher speech (using recording and process procedures described below). The measurement properties of the system are discussed in Section "Accuracies", and further in Dale et al. (2022). Its measurement properties were validated in secondary school English and language arts classrooms in western Pennsylvania, the same setting the present data were collected in. Although this close match between model development is likely to reduce or eliminate any concerns about bias in the present study, extensions to other contexts would be likely to degrade performance and potentially introduce bias without revision to design features. The system is based on a formative/additive conception of instruction, such that evidence from the full set of measures collectively constitutes or characterizes

discourse practices, and teachers may be stronger or weaker in specific areas. Definitions and examples of each discourse feature are provided in Table 1.

The three major steps towards providing automatic feedback include: (1) recording and transcribing teacher audio; (2) classifying discourse features; and (3) visualizing and feedback. The entire pipeline can run automatically; however, we included manual oversight (e.g., renaming mislabeled files, checking that audio was transcribed) at one key phase to ensure fidelity.

Recording & Transcribing Audio

Teachers were provided with a Samson 77 Airline headset microphone system and a laptop with recording software called RecordPad. Details of the microphone setup are provided in Jensen et al. (2020). Briefly, however, the microphone is a high-quality unidirectional microphone with cardioid pickup patterns—the microphone only picks up sound from one direction (i.e., the teacher) and is most sensitive to sounds from the front of the mic, thereby canceling background noise. The microphone is more complex than for example, recording to a smartphone, but it provides high-quality audio, which enhances accurate discourse classification (See D'Mello et al., 2015). Teachers were trained on how to use the system in an instructional session and our previous results (Jensen et al., 2020) indicate that they can effectively use it to independently record high-quality audio.

Recordings were saved to a Dropbox folder set up to automatically synchronize audio with cloud storage, which was then available to our processing pipeline. Once the recording was detected on the cloud, an automatic processing pipeline commenced. After completing basic quality checks (e.g., was the file successfully transcribed), the audio was submitted to the IBM Watson automatic speech recognition web-based service, which provided transcriptions of teacher utterances along with start and end times for each utterance. To alleviate occasional inaccuracies in automatic segmentation of individual utterances, consecutive utterances with less that 1-s of inter-utterance duration were merged. The microphone was highly accurate at filtering out student speech, but student speech does occasionally bleed through. This was addressed by running all transcribed utterances through a teacher vs. student speech classifier (similar to Section "Discourse Feature Classifications"), which leverages inherent differences in speaking pattern to filter out student speech with very high ($r=0.93$) accuracy. Previous work on similarly collected data using IBM Watson indicated a transcription accuracy of 72% (Jensen et al., 2020). Critically, correlations between speech recognition errors and the accuracy of teacher talk classification were quite low (-0.03 to 0.15) indicating considerable robustness to these errors.

Discourse Feature Classifications

We used a modern deep (machine) learning architecture called *transformers* (Vaswani et al., 2017) to automatically code each utterance for evidence of the five discourse features. The machine learning is based on gold-standard human coding, where experts in ELA instruction coded the presence or absence of each feature (the vast majority of codes are simple binary codes) at the utterance level, which are

then aggregated to produce observation or lesson-level prevalence rates (all observations are of full, ~ 50-min lessons. See Dale et al. (2022) for further details). This is an adaptation of the coding approach carried forward from the Nystrand & Gamoran era,[4] with two updates to coordinate with our automated methods; first, coders worked from transcripts generated by automatic speech recognition and automatic utterance segmentation. Second, coders coded all teacher utterances (our system focuses only on teacher speech), unlike in the Nystrand and Gamoran era when only teacher questions were coded. These methods are more fully described in Dale et al. (2022).

Our machine learning approach is a *transfer* method, where a model trained on one dataset/task is adapted to another (Pan & Yang, 2010). This entails two steps: *pretraining* and *fine-tuning*. During pretraining the transformer uses large amounts (i.e., gigabytes) of text to learn the meanings of words specific to their context. The resultant *contextual representations* of words, serves as a starting point for subsequent *fine-tuning*. Here, the computational model is then fine-tuned (parameters are updated) on small amounts of task-specific data (teacher discourse classification in our case).

The specific transformer model we used is called bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018), which was pretrained on the entirety of English Wikipedia (2.5 billion words) and a corpus of 11,308 unpublished books on the web (800 million words). We used the HuggingFace's transformers (Wolf et al., 2019) library's implementation of the BertForSequenceClassification model and the BertTokenizer and fine-tuned the BERT model for two epochs using a batch size of 32.

Fine-tuning was done using a previously collected set of 16,977 teacher utterances from 16 teachers across 127 classroom audio recordings using the same microphone and automatic transcription method used here (Jensen et al., 2021). The utterances were annotated by trained coders for the various discourse features; the coders achieved an average reliability of 0.81 (Gwet's AC; Gwet (2008)), which we deemed sufficient for automation. Utterances were individually inputted to the model. Critically, the training procedures ensured that the models could generalize to data from new teachers (i.e., teachers not represented in the training data).[5]

The output of the computational models is a set of probability distributions across the target discourse moves, one per utterance, which were then averaged across all the utterances in the recording. These are then converted into ordinal categories

[4] The first author participated in the Nystrand and Gamoran studies beginning with the national or five-state study (Gamoran & Kelly, 2003), and then the Partnership for Literacy Study (Kelly, 2008).

[5] Models were trained and evaluated using tenfold teacher-level cross-validation, where all utterances for a given teacher were either in the training set or the testing set, but never in both. If a teacher in the present study also contributed data used to train the models (prior data collection), the models were retrained after removing utterances from that teacher prior to use in the current study. In this fashion, there was no data overlap between model development (training) and deployment.

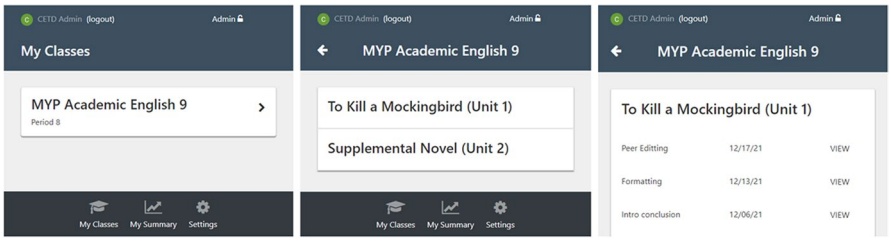

Fig. 2 Teacher Talk Tool organization by classes, units, and lessons

(low, medium, and high) based on the 33rd (low) and 66th (high) percentile values based on the above training sample of 127 recordings.[6]

Accuracies It is difficult to identify a single standard of successful automation as measurement error in automated systems exists in a trade-off with efficiency; a given analyst may be willing to sacrifice more or less fidelity of measurement in exchange for much larger sample sizes. In studies comparing the automated scores with gold-standard human codes (Jensen et al., 2021), these models have average utterance-level accuracies measured via the area under the receiving operating curve (AUROC) of 0.84 (0.83 for Instructional Talk, 0.73 for Open-Ended Questions, 0.86 for Elaborated Feedback, 0.90 for ELA Terms, and 0.88 for Goal Clarity). When averaged to the observation/lesson level, this corresponds to an average correlation of 0.57 (range of 0.35 to 0.70), and average percent agreement using tercile cut points of 52% (range of 0.42 to 0.66). Together, these exceed that obtained in large scale studies of human observations of classroom practice (Ho & Kane, 2013; Kelly et al., 2020a, b).

Front-End Visualization Web Application

The web-interface (for both mobile and desktop web) helps teachers navigate their recordings and visualize the automated feedback for individual recordings and across sessions. The main navigational features shown on the bottom menu include: My Classes, My Summary, and Settings.

Classes, Units, & Lessons Pages Teacher feedback within the app is organized into the following hierarchy: Classes > Units > Lessons. Thus, navigation through the web application follows that same pattern. A teacher may click into one of their classes to view their units within that class, click into a unit to see the lessons within that unit, and click into a lesson to see lesson-specific feedback. Figure 2 contains, from left to right, screenshots of the classes, units, and lesson pages, respectively.

[6] The first participant received labels based on 15–85 cut points, which we quickly realized obfuscated far too much important variation in talk features.

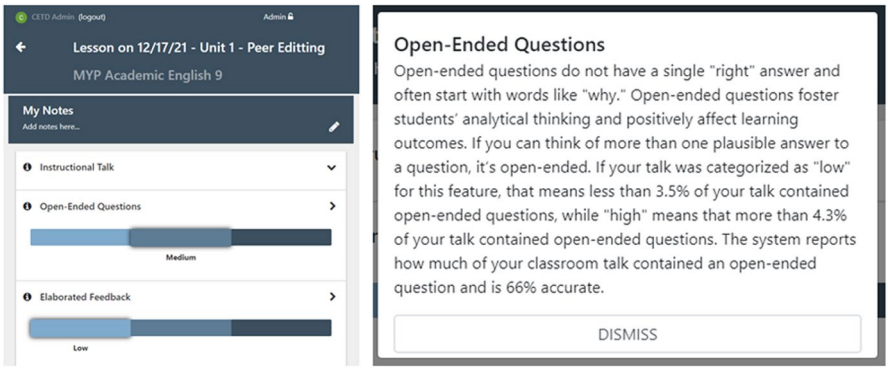

Fig. 3 Lesson level feedback page (partial view)

Lesson-Level Feedback Page When a teacher selects the VIEW option for a given lesson in the Lesson page, a lesson-level breakdown of feedback is presented for each discourse feature. The feedback for each feature is presented in ordinal categories of either Low, Medium, or High (See Fig. 3). Teachers can click the information *i* icon to the left of the discourse feature name to view a brief description about the discourse feature, the cutoffs used to determine the ordinal category, and the accuracy of the automated scoring for that feature. There is also a My Notes section along the top of each lesson page in which teachers can record comments and thoughts.

Summary Page Finally, teachers can also view feedback across all lessons within the My Summary page. The page is a simple breakdown of the proportion of lessons that fall into the Low, Medium, and High categories of feedback for each of the five discourse features (See Fig. 4). For example, in Fig. 4, 13% of the teacher's lessons received a "high" on the instructional talk category. Teachers can click on the *i* icon to the right of an ordinal category name to see more information about the summary within that category.

Participants & Lessons

The study setting was a large, predominantly white, High-SES public high school in Western Pennsylvania. The study site was selected because it is a member of the Tri-State Area School Study Council, a professional development organization partnering with the study team. Participants included four white female and one male English teacher. All teachers had master's degrees, full, regular state certification, and entered teaching through a traditional teacher education program. Participants' years of experience ranged from 10 to 24 years. All teachers reported high levels of efficacy at the start of data collection on a set of nine items about the functions of talk in ELA classrooms (e.g., "To what extent do you feel successful in leading classroom talk to support students' understanding of a text's central idea or argument

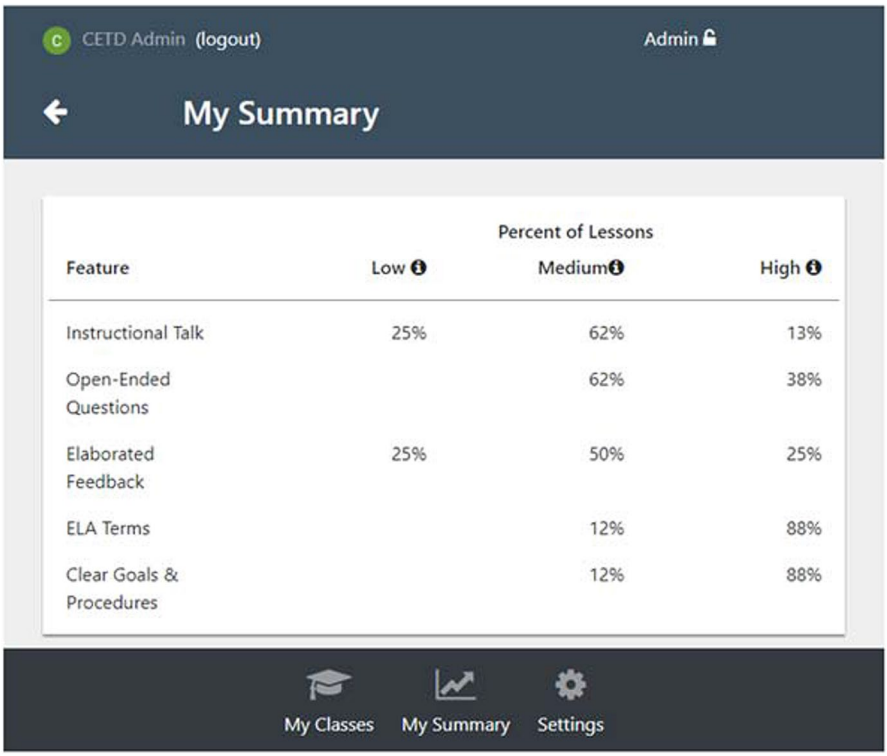

Fig. 4 Summary page

(including claims, evidence and reasoning)?").[7] Table 2 lists each respondent, the grade and achievement level of the focal class recordings were carried out in, and the instructional units of the recorded lessons. Class sizes ranged from 20–25 and included students with a modal age of 14–15 (grade 9) to 17–18 (Grade 12). Classes were general education classrooms containing primarily English proficient students ($\leq$2 English Language Learners per class), and entirely (as reported by teachers) non-poor students.

Recording and Feedback Procedure

Teachers independently recorded eight lessons, scheduled at their convenience on days when the teacher was interacting substantially with students in whole-class or

[7] Efficacy items were newly developed for this study based on learning/standards goals from the Common Core State Standards for English language arts and Literacy. Although the items appear highly internally consistent (Cronbach's alpha of above .9), various survey response processes (e.g., adjacency effects) can artificially inflate such statistics. The mean of the efficacy items was 3.33 at the start of data collection (on a 4-point scale), increasing to 3.64 at the end of the study.

Table 2 Participants and their focal classes

Participant Pseudonym	Grade level and achievement of English class		Main content of lessons
Roxanne	11	On grade level	*The Great Gatsby*; review and analysis (4 lessons), *The Raven*; background, analysis (2 lessons). *The Crucible*; historical background, review and analysis (2 lessons)
James	9	On grade level	Writing project on *To Kill a Mockingbird*; quote incorporation, formatting, vocabulary, outlining, thesis statement, peer review (7 lessons), final discussion. Supplemental novel (1 lesson)
Annie	9	On grade level	Writing project on *To kill a Mockingbird*; outlining, evidence search, thesis statements, quote incorporation, etc., and peer workshop (8 lessons)
Erin	9	Above grade level	*Romeo and Juliet*; cultural/historical background, annotation, review/analysis and discussion, performance (8 lessons)
Mary Anne	12	Above grade level (AP class)	Writing project on justice in *Born a Crime* and Solomon's "Justice and a Passion for Vengence;" introduction, constructing an argument essay, rhetorical analysis, quotation analysis, revising, discussion, + review of approaching AP question (part of 1 lesson)

supervised monitoring of small group or individual work. Before receiving feedback after their third lesson (the first three lessons were recording-only), teachers were given a short PowerPoint presentation (via zoom) on the Teacher Talk Tool. This presentation emphasized the importance of classroom communication to student engagement and learning and provided a substantive overview of talk features. Additionally, several elements of the tool were stressed in this presentation. First, we stated that the Teacher Talk Tool does not *evaluate* their talk, although the presence/feeling of evaluation was not fully absent (see discussion section). Second, throughout the presentation (and in the information available in the tool itself), we stressed that the feedback is not fully accurate. For example, for goal clarity we stated, "The system determines how much of your classroom talk helped identify goals and explain procedures and is 68% accurate."[8] The accuracy of the feedback for a given observation is in fact unknown (as reliability and validity estimates are aggregate properties of the system). More importantly though, users should know that the

[8] Accuracies were reported like this in two places: in the initial overview presentation to teachers, and in information screens on the Webapp. When the system was switched to tercile cut points for low, medium, and high scores, we failed to correctly update the cutoff-based accuracies reported to users (i.e., we continued to use the accuracies corresponding to original 15/85 cut points). This error notwithstanding, we are confident users were well-apprised that the system is not fully accurate.

underlying information has errors and that it is always possible a given classification is inaccurate. Counterintuitively, by helping prepare users to see unexpected feedback, this may improve overall trust in the system.

After the 4th lesson, teachers began receiving automated feedback through the app. Specifically, teachers received email notification that their lesson was successfully recorded and processed, and feedback was available for review through the web application. Several basic quality features were reviewed by a study team member before notifying participants, generating a lag of several hours, although the system itself processes a 50-min lesson in approximately 40–60 min. This delay was by design and the tool was never intended to provide feedback on a recording immediately after it occurred (impractical for busy teachers), but for review at a later time. Over the course of the study 15 errors occurred in processing that required manual intervention (e.g., renaming the file), the majority of which were user errors in file naming.

Feedback Received

Table 3 summarizes the feedback scores participants received as averages of the low (1), middle (2), and high (3) categories across eight lessons. We note mean scores ranging from a low of 1.33 for instructional talk, to a high of 2.40 for goal clarity. Instructional talk (mean of 1.33) proved to be a problematic construct. The variation in the underlying proportion of questions and statements that are instructional (as opposed to non-academic talk about home and school matters) is low, such that very small differences in instructional talk led to changes in classification. At the same time, the overwhelming majority of talk is instructional (77.2% in this study). Thus, the vast majority of the lessons scored "low" in instructional talk in this study were neither low in any meaningful absolute sense, or even very different from a lesson that would have been scored as "medium." While this construct might be useful in another research context, it was a dud or a "lemon" here.

Yet, "making lemonade" out of the virtually useless feedback on instructional talk, consider that it may nevertheless have served a positive rhetorical function in this feedback-response study, because it might have served to balance the otherwise positive scores. That is, if receiving positive scores induces participants to report more positively about accuracy and other aspects of receiving feedback (positivity bias), then the presence of a typically lower score in instructional talk may have countered that positivity bias somewhat. At present, we would not recommend including instructional talk as a feedback construct in a widely used system, but it may have served a useful role in this study. Table 3 also reports an informal measure of relative variability across teachers (compared to within teachers), the Intra Class Correlation coefficient (ICC).[9]

[9] We designate this statistic as informal because the ICC in such small samples is readily impacted by chance/random differences across teachers.

Table 3 Feedback received on each talk feature: Mean of low (1), middle (2), and high (3) across eight lessons

Participant	Instructional Talk	Open-Ended Questions	Elaborated Feedback	ELA Terms	Goal Clarity
Roxanne[a]	1.25	2	1.75	1.88	2
James	1.88	2.38	2	2.88	2.88
Annie	1.25	2	1.38	2.5	3
Erin	1.25	2	2.5	1.88	1.63
Mary Anne	1	2.25	2.5	2.25	2.5
Grand Mean:	1.33	2.13	2.03	2.28	2.40
ICC[b]:	0.27	0.09	0.33	0.29	0.70

[a] Roxanne received 15–85 split feedback; middle scores designated/included underlying prevalence rates in the 15th to 85th percentile

[b] We present the ICC (from STATA's loneway command) here as an *informal* measure of the proportion of variance in scores that lies between teachers (compared to within teachers). With this sample size, estimates are unstable (with high standard errors). The ICC for open-ended questions is actually truncated by STATA to 0, but clearly there were differences between teachers and we report 0.09, which is the mid-point of the 95% CI for the ICC

Survey and Interview Instruments

Multiple survey and interview instruments were used throughout the data collection (see Online Appendix). First, paper and pencil teacher surveys were administered prior to the recording sequence and at the conclusion of the study. The pre-study teacher survey included measures of socio-demographic and professional background and a 10-item Likert-scale battery measuring discourse-related efficacy. The post-study teacher survey: repeated the discourse efficacy items; included eight Likert-scale items asking teachers to reflect on dimensions of feedback throughout the study; and concluded with a classroom roster providing summary information on student background.

Additionally, two sets of short, Internet-based repeated surveys were used to assess teachers' feedback response; post-recording surveys completed as soon as possible after the lesson, and post-feedback surveys completed after receiving feedback on the app. Given the overall intensity of the data collection experience for teachers, our goal was to keep these repeated surveys as brief as possible to maintain teacher concentration and goodwill. Perhaps as a result, there was almost no item missing data. The post-recording survey (n = 39, 1 unit missing) featured a mix of 11 open and closed-, Likert-scale response items, while the post-feedback survey (n = 24, 1 unit missing) contained five open and closed-response items. Finally, online video-conference interviews were conducted by the first author after the third lesson, and again after the 8th lesson. The interviews were scripted, but also interactive with the participants looking at feedback from the app in real-time, lasting approximately 50 min. Interview items were designed to measure reflection (attention & sense-making) and feedback uptake (goal setting and strategy use), along with additional checks for perceptions of accuracy.

Collectively, the surveys and interviews measured a set of conceptually interrelated and consequentially valid outcomes similar to those used in prior research (Brett & Atwater, 2001; Chawla et al., 2019; Quintelier et al., 2020) Perceptions of accuracy and importance, teachers' attention and sense-making, and feedback uptake were assessed from multiple sources, including the post-feedback survey and each of the two interviews. For example, the post-feedback survey asked teachers: "As a whole, how accurate do you feel the Teacher Talk feedback was for this class session? (on a 4-point scale; not at all accurate, only somewhat accurate, mostly accurate, highly accurate). Additionally, throughout the interviews, teachers were asked "Do you think this feedback is accurate" when viewing results for individual talk features in individual lessons and on the My Summary screen. Perceptions of importance (whether a given construct plays an important role in instructional processes and outcomes) were assessed primarily from interview data, when teachers engaged in sense-making about the sources and determinants of talk features. Attention-focusing and sense-making was assessed with closed-items and throughout the interviews as participants were asked which particular talk features stood out to them, whether they were satisfied with their results, whether the feedback matched their impressions, and how the feedback aligned with how they typically think about their classroom talk. Feedback uptake was assessed with survey and interview items that all required teachers to compose/report goals, etc., to reduce acquiescence bias.

Analytic Methods

Analytic methods include a mixture of quantification of repeated-measure survey data, coding and quantification of accuracy responses in the face-to-face interview data and further qualitative appraisal of the interview data including a case study of one participant. Interviews were conducted by the first author, and then coded and analyzed by the second author. While the sample size precludes formal tests of statistical significance in the survey data, the surveys were a valuable complement to the interview data, providing consistency and standardization in querying teachers about key constructs. Overall, we focus on a basic assessment of central tendency and variability, and congruence across data sources, in understanding teachers' feedback response.

To begin the interview coding process, each interview was converted into transcripts via otter.ai.[10] Since each of five teachers were interviewed twice, with content overlap across the first and second interviews, we chose to code and analyze both interviews for each teacher sequentially before moving on to the next teacher. Interview data was coded using a combination of deductive and inductive approaches in multi-cycle coding following Vanover et al. (2021). The primary coding was conducted by the second author with auditing by the first author to achieve a consensus,

[10] Quotes included in results here have been corrected for word substitutions and various errors.

strengthening the validity and trustworthiness of our findings (Miles et al., 2020). In the following results section we integrate findings from the survey and interview data to reach an overall inference about the dimensions of feedback response.

Results

RQ1: Accuracy & Importance

Responses on the post-feedback survey indicated high levels of perceived accuracy (Table 4). In repeated survey responses following lessons 5–8, the majority of feedback was judged to be mostly (54%) or highly (25%) accurate, with only a minority of lessons judged to be only somewhat accurate (17%) or not at all accurate (4%). Respondents were also asked specifically at several points in the interview (twenty times in all across all interviews) when looking at both individual lesson feedback and summary feedback whether they thought the feedback was accurate. When asked directly about accuracy in the interview, 40% of responses clearly affirmed accuracy, and the remaining 60% partially affirmed accuracy (e.g., "If I had to guess, I would say that maybe it's 80% accurate…"), with no clearly negative appraisals of accuracy. The face-to-face nature of the interview may have biased these responses (where negative appraisals were withheld to please the interviewer), but *both* the survey and interview data convey a positive appraisal of accuracy.

Throughout the study, feedback on the instructional talk construct was a source of confusion for respondents, negatively affecting perceptions of accuracy and serving as a source of distraction. Scores on instructional talk were far lower than any other feature. Overall, this construct functioned incompletely within our conceptual model. It did produce a very strong attention-focusing effect as the feature with consistently the lowest scores. Yet, it lacked face validity to respondents, and we realized mid-study that this construct was not a good candidate for individual-level teacher feedback. This construct was included in the feedback system in part because it is essential to the coding framework the system is validated on, where utterances are labeled as instructional/non-instructional to structure coding. Yet, in the reference corpus, which determines the feedback ratings, the distribution of the instructional talk measure is quite compressed with a majority of talk labeled as instructional; the interquartile range for instructional talk is 0.796–0.914 (Dale et al., 2022). Thus, in the typical teacher's classroom, we would not expect much substantively meaningful variation in this measure.[11]

It appeared that in the teachers' view, all of their talk was "instructional," and they consistently and repeatedly wondered what this construct was referring to (even as it was defined for them with a similar level of detail to the other constructs). In other

[11] The instructional talk measure could be very useful in other contexts, such as making more highly aggregated appraisals (e.g., across schools or districts), or in larger scale studies where the tails of the distribution would be relevant.

Table 4 Survey evidence of perceptions of accuracy, importance, attention focusing, and feedback uptake

Item [construct][a, b]	Source	Responses: Frequencies (percentages)			
		Not at all accurate	**Only somewhat accurate**	**Mostly accurate**	**Highly Accurate**
How accurate was feedback for this lesson? [A]	Post-feedback survey	1 (4%)	4 (17%)	13 (54%)	6 (25%)
		No	**Yes, somewhat**	**Yes, Substantially**	
App highlights crucial talk? [I]	Teacher survey	0	2 (40%)	3 (60%)	
App helps focus attention? [F]	Teacher survey	0	4 (80%)	1 (20%)	
App feedback prompts me to consider changes? [U]	Teacher survey	0	4 (80%)	1 (20%)	
		Focused about the usual amount	**Somewhat more focused on classroom communication**	**Much more focused on classroom communication**	
Did conducting a recording make you more focused on classroom communication? [F]	Post-recording survey	22 (56%)	16 (41%)	1 (3%)	
		No specific goals/strategies		**Goals/strategies listed by teacher**	
Do you have specific goals for next lesson? [U]	Post-feedback survey	10 (42%)		14 (58%)	
Do you have specific changes/strategies planned for next lesson? [U]	Post-feedback survey	15 (62%)		9 (38%)	

cases they mistakenly equated this feature with "direct instruction." Discussing accuracy, Erin (all names are pseudonyms) stated:

> "Based on what I know, like what I recorded, it probably is [accurate], except for that instructional talk; that's bothering me, driving me crazy."

And later,

> "…If I really am asking a lot of open-ended questions and giving elaborated feedback, and I'm using ELA terms, like why isn't that instructional [talk]… why is that so low? That's really the one that I feel like is not accurate to me."[12]

To summarize, respondents reported positive appraisals of accuracy, with the exception of the instructional talk construct, or we would even say, *in spite of* receiving feedback on a construct with such low face validity.

Based on focus groups conducted prior to this study, we suspected that teachers would find the constructs measured in the Teacher Talk Tool important and relevant to their instruction. When asked on the post-data collection survey whether the results highlighted an area(s) of classroom talk that was crucial to the success of their instruction, all participants responded affirmatively (Table 4), either "yes, somewhat" (Annie, Mary Anne), or "yes, substantially" (Roxanne, James, Erin). In interview data, importance and relevance was also indicated obliquely by reference to what teachers themselves perceive to be their own strengths and focus, as in the following quote:

> "Yeah, I mean, I think that having very clear expectations is, I would consider that one of my strengths." (Annie)

RQ2: Attention and Sense-Making

To assess attention focusing, we began by considering the constructs that teachers focused the most on in terms of their a priori attention (i.e., that they mentioned as most important to them before using the tool) and then, how that appeared to change upon receiving feedback. Roxanne did not articulate a particular a priori focus, but James appeared to be less focused coming into the study on open-ended questions and elaborated feedback. Annie barely mentioned open-ended questions throughout her interviews. Among the constructs, Erin paid the least attention to goal clarity (only mentioned once, in passing). The following quote from James illustrates his a priori emphases and the focusing effect of the app feedback:

> "Well, again, I think pretty consciously of the goals and procedures, so it's good to see that on there. And it's good for my ego to see the ELA terms so high. So, I guess, the goals and procedures, I'm going to stick to that, that's

[12] This quote illustrates the depth of Erin's puzzlement over the instructional talk feature but also fundamental misunderstanding of the features; instructional talk is estimated, by definition, orthogonally to the other features. The inter-relationships among features is understandably challenging for a new user to understand.

generally how I think in my lessons. The rest is something that maybe I didn't think of as consciously until I looked at the app." (James)

In the interview data, it is difficult to fully separate out teachers' a priori focus from focus elicited by receiving app feedback, even if they use a past imperfect phrasing, because they had already received feedback at that point. Survey data provide a more direct look at attention focusing. On the post-data collection survey all respondents agreed the app helped focus attention somewhat (Roxanne, James, Annie, Mary Anne) or substantially (Erin) on teaching strategies or techniques they would like to improve (Table 4). In the post-recording survey, participants were asked whether conducting a recording alone (prior to receiving any feedback) made them more focused on communication. Forty-four percent of participants reported being somewhat or much more focused on classroom communication, while the remaining 56% of recordings they were focused "the usual amount." Relatedly, various sense-making efforts in the interviews by teachers revealed basic attention to specific constructs along with cognitive processing activated by receiving feedback. In making sense of app feedback, teachers referenced their goals, their efficacy (what talk features they thought of as their strengths), as well as lesson content, and to a lesser extent student readiness:

"Again, you know, clear goals and procedures is interesting. That one was interesting to me, because I think if I had recorded my academic classes, those numbers would have been completely shifted, because I have to spend a lot more time explaining and guiding and walking things, walking them through things. Whereas with my honors kids, I could just throw it at them super quick and they like pick it up. They get frustrated if I over-explain what's going on because they get it. So yeah, I think that's, I noticed those things. It makes me think about the ways, how I talk differently to different groups of students. Even though I know that's not what this, this was just evaluating one class, but it makes me think about that." (Erin)

In summary, we generally found that use of the app (both in the recording itself and in receiving feedback) helped focus attention on talk features teachers deemed important, and that feedback elicited various forms of sense-making. One of the most important findings here is the pronounced teacher-to-teacher variation in *which* constructs teachers focused attention on. The system explicitly (in our instructions) and implicitly (by withholding any summary score that aggregated scores from all features) offered participants choice in which constructs to focus on, and the participants seized that opportunity.

RQ3: Feedback Uptake

To a certain extent, survey and interview data on sense-making provides a preliminary look at feedback uptake. Here, we consider more direct measures, including goal setting and strategy use. On the post-feedback survey, respondents were asked what goals they had, if any, related to teacher talk for their next lesson. On the repeated post-feedback surveys, participants responded affirmatively and listed

specific goals 54% of the time (Table 4). These goals directly referenced the talk features contained in the app, with one exception, where a participant planned to provide a "live example" (presumably addressing goal clarity, although that was not explicitly stated). Similarly, teachers were asked on the post-feedback survey what strategies, tools, or preparation if any, they would like to implement to support talking to learn in your next lesson? A total of 38% of respondents listed content, although in the majority of cases they again listed goals rather than actual strategies (i.e., "by doing…," or "do more of A and less of B"). The post-data collection survey asked a summary question, whether reviewing and comparing lesson results with the app prompted participants to consider a change in lesson plans or practices? All participants responded affirmatively (Table 4), but the majority selected "yes, somewhat" (Roxanne, James, Annie, Mary Anne) rather than "yes substantially" (Erin).

Interviews offered greater opportunity for teachers to discuss these "furthest reaches" of teacher learning. Roxanne received some of the lowest scores among participants, and overwhelmingly medium scores, so we would perhaps have expected her to be somewhat skeptical towards feedback. Yet, there was evidence of attention focusing and sense-making throughout her interviews. Feedback uptake was more limited but still present. She concluded:

> "I definitely think that I have been working toward more open-ended questions and elaborated feedback, probably since using this tool."

However, she did not discuss any specific strategies to achieve those goals. James received some of the highest feedback scores, but his use of the app definitely still seemed to elicit attention to classroom discourse. James also discussed explicit goals based on the app feedback, tying those goals to particular content:

> "I think, particularly during, particularly with the lecture lessons, and there are quite a few of those, maybe try to incorporate more of the questions and elaborated feedback into it, if I can, you know, I don't want to put it in artificially, but where I can put it in. Because looking at the app is funny. I mean, just by putting it here as a topic, it makes me think about it. So suddenly, you know, I think well, that's important, I need more open ended and more elaborated feedback. So, I'd like to maybe incorporate that into some of the more lecture-centered lessons."
>
> "Generally, so upcoming lessons we're doing work in vocabulary, and we're doing work in hopefully some poetry work coming up next. So, there will be a lot of opportunity for definitely, with the poetry, for example, the ELA terms, and instructional talk, hopefully. So, I guess those are the plans in the immediate future. And then down the road, we're doing some nonfiction stuff."

Annie was one of the participants whose lessons occurred during a writing project, and one of the topics that came up in her interviews was the nature of elaborated feedback in that context. Her scores were low on that feature, which prompted her to think about that construct. When asked about goals for the future, she discussed feedback at length, considering alternate strategies, including forms of

feedback from her (e.g., how much of that would occur as whole-class instruction), but also student-to-student peer feedback. Her plans were not fully formed, but she clearly had been prompted by the feedback to weigh options and adapt as the lesson developed. She also discussed goals for discussion-based learning:

> "Well, I think maybe to make it a little bit more student centered and more of a variety of options and opportunities for them to get involved in the discussion, instead of there being sort of a lot more of a right or wrong answer, more of like a formula to follow with their first multi paragraph essay, something that's a little bit more creative or more choice involved, that would give more opportunities for a deeper level discussion, I think with elaborated feedback and open ended questions."

In her survey responses, Erin indicated positive response to the app feedback in focusing attention and in feedback uptake. Her interview also indicated much sense-making, but unfortunately, she spent a great deal of time examining and discussing the results for instructional talk. In the end, her interview data did not demonstrate goal setting or strategy use with notable/sufficient depth, which contrasts with her survey responses. Thus, while there was a clear attention-focusing effect, we are not sure what to make of Erin's feedback uptake. Summarizing findings on feedback uptake from the survey and interview data, we find positive but quite limited evidence of feedback uptake in the form of goal setting and strategy use.

To examine the relationship between participants' focus and changes in observed practice, we compared the foci/goals expressed in interview one, and subsequent trends in the underlying scores (prevalence rates/proportion of discourse moves). Roxanne focused on improving elaborated feedback. Subsequent lessons showed a mix of high (above 60th percentile) and low (below 20th percentile scores). Annie focused on ELA terms and goal clarity, but with the exception of Lesson 2, Annie scored very high (above 80th percentile) on both of those constructs throughout her observations. James focused on elaborated feedback, with subsequent lessons showing generally low levels, below the 50th percentile, and very low ($<$ 10th percentile) in Lessons Six and Seven. Erin focused on open-ended questions, and with the exception of Lesson One where she scored very high, her scores generally improved in later lessons. Of Mary Anne's three foci, there was a mix of outcomes with no clear trends. Overall, how well participants were able to actively improve scores is difficult to discern due to the strong effects of lesson context on discourse practices.

Case Study of Mary Anne's Response to Feedback

Mary Anne's Background and Focal Class

Mary Anne was the most experienced teacher among the participants with 24 years of experience. Mary Anne entered teaching through a traditional teacher education program. She held a Master's degree as well as an educational specialist diploma, and was fully certified by the state, but did not held, and had never sought NBPTS certification. Forty-five years old at the time of the study, she had been teaching at

her current school for 21 years and served as chair of the English and Language Arts department. She had not participated in any professional development activities related to classroom talk in the past year. At the start of the study, we asked Mary Anne about her overall ability to lead/manage classroom talk and discussions. Experienced and confident, Mary Anne reported feeling highly successful (or efficacious) at leading classroom talk for a variety of purposes, including activities with a substantial writing component, but she did indicate she was just "moderately successful" at leading classroom talk in some areas, including: analyzing texts to learn about author's craft and structure, analysis of different mediums (e.g., drama, multimedia), narrative story-telling, as well as one of the major learning goals in ELA, understanding of a text's central idea or argument.

Mary Anne's focal class was her 12th grade Advanced Placement (AP) English class, which she reported as containing 17 white students, 4 Asian students, and 4 other/multiracial students, the majority of whom were achieving "above grade level." When Mary Anne's observations began, the class was just completing a research paper, and the first observation focused on elements of revision and editing. The second observation occurred on a day when the class was reviewing multiple choice question strategies for the AP, and the upcoming Justice unit was introduced, which would include the non-fiction philosophical text "A Passion for Vengeance" by Robert Solomon (1995) as well as the memoir *Born a Crime* by Trevor Noah (2016). Subsequent lessons in the unit on justice in literature included identifying elements of the rhetorical situation (the structure and appeals used by the author) in Lessons Three and Four; author's use of sentence variety, structure, figurative language, comparison and other elements (Lesson Five), analyzing key rhetorical strategies in specific passages (Lesson Six); and analysis of key quotations, where students interpreted quotes, took positions, and offered supporting evidence from the text (Lesson Seven). Finally, in Lesson Eight, the class transitioned to how to approach a cold (never before seen) prompt and begin an argument essay. Over the course of the lessons, the class utilized a wide variety of activity structures, including teacher lectures, individual seat work, work in small groups, and less commonly, whole-class discussions (Lesson Four). In the lessons Mary Anne submitted for feedback, she reported the students were generally engaged (e.g., highly interested, enjoying class a great deal), although there was some variation from lesson to lesson, and she reported only modest levels of student concentration compared to other dimensions of engagement.

Mary Anne's Experience with Receiving Feedback from the Teacher Talk Tool

Mary Anne reported that the Teacher Talk Tool system was generally easy to use. She reported that the feedback was timely, she was able to readily find information and explanations in the app, and the feedback on the lesson and summary pages contained the right amount of information for her. Over the course of her lesson observations, Mary Anne would generally receive medium and high scores. Apart from instructional talk (see previous discussion), she only received one low score, for open-ended questions in Lesson Eight. Each lesson contained a

mix of medium and high scores. She received high scores for elaborated feedback in observations 1–4, and then medium scores thereafter. Her scores for open-ended questions were lower in lessons 1–2 and 6–7, while her scores for goal clarity were higher in lessons 1–2 and 7–8. While she experienced slightly lower overall scores in later observations, there was not a clear or strong discernable trend in her scores over the course of the eight lessons.

As Mary Anne made sense of the feedback she received in the two interviews she framed her interpretations in two ways, with reference to what she sees as her pedagogical strengths, and how different instructional goals, activities, and content would logically elicit different discourse dimensions/emphases. In the first interview, Mary Anne explained:

> "I work really hard to ask open ended questions of my kids and…, I don't want this to sound like arrogant in any way, but I think that the good teachers give elaborated feedback. Just saying, "great," isn't really helpful to kids. Kids learn more when you elaborate, so that the entire class can hear why it was right or how it maybe needs to be tweaked. So, I think that that those two certainly reflect my efforts and where I like to set my goals in my teaching."

Later, Mary Anne also highlighted goal clarity:

> "I guess those three categories are the ones that as a teacher are most important to me: open ended questions, elaborated feedback, and clear goals and procedures."

Yet, in contrast to her a priori expectations, Mary Anne did not receive the highest score consistently in any of the domains. Reflecting back on her overall feedback in Interview 2, she expressed surprise that her feedback was not more consistently high than medium. Looking at specific lessons, Mary Anne readily discussed how features of the lesson likely influenced her discourse feedback. In some cases, that interpretation involved locating what she viewed as a logical source of a lower score:

> "In this particular lesson, we had a whole class discussion, and then I broke them into groups. And so, I think, because of the level of my students, and because of, you know, where we are in the year that, they didn't quite need, like, minute directions." [referencing a medium score for goal clarity]

In other cases, the feedback proved validating, because she felt challenged by the material itself:

> "The piece that the kids read for that day was very challenging. It was philosophical in nature and a little bit esoteric, so, I was doing my best to kind of like draw them out…by elaborating, and validating, and redirecting when necessary, because it was hard. It was really hard for the kids. And I teach, you know, some of the brightest kids in the school. So, I guess I'm pleasantly surprised." [reflecting on the high score for open-ended questions]

Reflecting on the overall utility of the system, Mary Anne categorized the feedback in survey responses as *somewhat useful* in highlighting areas of classroom talk

that are crucial to the success of her instruction, in helping focus attention on teaching strategies or techniques she would like to improve, in prompting her to consider possible changes in lesson plans and practices. As the above quotes indicate though, as maybe we should expect from a highly trained, veteran ELA teacher, as she processed her feedback, Mary Anne actually engaged in a "close reading" of the feedback. Her attention was clearly focused on the talk constructs, and she engaged in a variety of sense-making analyses, comparing the feedback to her own expectations, and contextualizing the feedback within the lesson content and goals.

Although Mary Anne almost never received the lowest category of scores, she often treated medium scores as *low to her*, and so such feedback might have prompted uptake in the form of goal setting and strategy use. Her survey responses indicated some focus on bolstering elaborated feedback in her classes. We also saw some glimpses of goal setting in the interview data:

> "It might give me pause to think about my goals and procedures. I think it's really important that kids are clear on what you're asking them to do. And it might give me pause, just to think, was I not clear enough? Do I need to maybe be more conscious of that moving forward?"
> "I think that I would be more cognizant of open-ended questions…"

Yet, in Mary Anne's experience with the feedback system, whereas there was much evidence of attention-focusing and sense-making, we didn't see robust goal setting, or any specific strategy use.

Discussion

To date, the vast majority of research on instructional observation has concerned the basic measurement properties of observational protocols, rather than their usage—how teachers respond to and learn from the process of observation (e.g., Quintelier et al., 2020). This response likely depends substantially on basic features of the observational process, including the underlying purpose of the observation (i.e., for evaluative vs. developmental purposes), and who is doing the observing (see e.g., studies of peer feedback by Wylie & Lyon, 2020). Systems relying on automated feedback naturally raise questions of whether teachers respond positively to this developing approach to feedback (Demszky, 2022; Jacobs et al., 2022; Korban et al., 2023). In this study we examined a set of interrelated research questions concerning teachers' perceptions of accuracy and importance, attention focusing, and uptake of feedback. Survey and interview data show support for our conceptual model of teacher discourse feedback. Participants understood the features of teacher talk reported by the app, as a whole, to be important and central to their instruction. Yet, importantly, teachers showed substantial variation in which features they focused on the most. We believe this may be a critical basic insight into how teachers use feedback systems, and that emphasizing choice may lead to greater user-engagement (Kelly, 2023). That

being said, the provision of choice was a constant rather than a variable in this study.

Another key feature of the system, which also did not vary but likely substantially affected the user experience, was that the feedback was not presented as [essentially] evaluative. This is true in several senses of design and presentation, but ultimately evaluation was not entirely withheld. In contrast to global observation protocols, no overall evaluation of the lesson as effective or ineffective is given in the Teacher Talk Tool, nor is teachers' use of particular talk moves deemed effective or ineffective. The tool is also more agnostic in its coding in that it tends to withhold judgment by avoiding categories such as "proficient", "deficient", etc. Relatedly, as discussed, we also stressed that teachers themselves should choose what to focus on, drawing on their own professional judgement, and that results are often affected by the lesson context, goals, etc. That positioning may have further reduced feelings of evaluation/judgement. Yet, this version of the tool was clearly comparative in providing scores of low, middle, and high. The features were also presented as being of broad relevance and importance. Results indicate that the users in fact often strived to obtain a higher classification, and viewed lower classifications negatively. As Mary Anne's case study indicates, even medium sores might be viewed negatively by confident, experienced teachers. Yet, there was also considerable evidence they filtered and appraised scores using their own professional judgement, and that as intended, the feedback was not viewed as immutably evaluative. Thus, in function, the tool was evaluative, but in a somewhat "softer" form than global protocols.

As one of the first studies to provide automated feedback, we were very interested in perceptions of accuracy (RQ1). Across multiple measures teachers perceived the app feedback to be generally accurate. In reflecting on feedback participants focused the most on constructs of a priori interest (as near as we can tell), and on constructs they felt were most relevant to their lesson content. Within that framework though, score-level seemed to have a very strong effect on teachers' attention. However, imbalance in attention to specific features was not as pronounced as in the Jacobs et al. (2022) study; each of our participants engaged with feedback from multiple talk features. One explanation for the pattern of findings in Jacobs et al. (2022) was that a single feature, the ratio of student to teacher talk, was so compelling and intuitive that it drew attention away from other features. Both very high, or very low face validity and/or perceived importance in specific features can absorb user attention.

Another overall finding that emerged was the consistently high degree of specificity that came through in teachers' reflection and sense-making (RQ2). The participants as a whole were not unlike Mary Anne; they "read' and analyzed the feedback they received, and processed it through the lens of their own context, goals, and understandings of their strengths and emphases. The level of abstraction of the features seemed to be a good match for teachers' own thinking, and they readily embraced the specific concepts and terminology used in the app. We believe this may be a basic principle of teacher feedback; if you present teachers with specific, concrete features, then this elicits teacher reflection at that level of specificity. One compelling way to elicit even greater reflection may be to provide access to examples

(e.g., audio clips, word clouds, etc.).[13] Overall, evidence of feedback uptake in the form of goal setting and strategy use was limited in this minimal-intervention study.

Like all studies, our has limitations. One pertains to the sample size of five teachers. Whereas it did feature a rich data set with 40 real-world classrooms sessions, 10 interviews, and several self-report questionnaires, the sample size precluded more formal statistical analyses beyond central tendency measures. A larger sample might have also have given much greater insight into feedback uptake, not only its prevalence, but it's possible forms and functions. A larger sample would have also allowed us to consider potentially key heterogeneity in response among novice and experienced teachers, and other forms of heterogeneity.

Another pertains to the lack of diversity in our teacher sample and relative homogeneity of the school we worked with. The present sample was very well matched to the development sample, so that may have promoted especially robust feedback. Yet, more basically, we are interested in how this approach to professional feedback might be adopted broadly, and "experimented" (in the Clarke & Hollingsworth, 2002 sense), with in a variety of contexts, uses, and settings. To that end, the study team has now partnered with the startup TeachFX to continue work in this field. This partnership will allow for broader work with teachers throughout the US and beyond.

Although the present study included repeated [eight] observations per teacher, a longer time frame might be needed for the effects of the feedback to be fully realized. For example, are the featured discourse constructs prevalent in teachers' thinking in the following academic year? Relatedly, our study was not experimental or otherwise comparative, where for example, we *varied* the type of feedback. Such designs could allow comparisons of perceived accuracy, etc., in fine-grained, agnostic systems like ours, with traditional in-person observation using global protocols. We would also be interested in studying even more basic, lower-inference features of discourse such as the ratio of student to teacher talk (a measure featured both in the Talk Moves and TeachFX systems). Having alternative feedback approaches in the same study, with identical dependent measures of feedback response, would provide a comparative frame of reference.

Lastly, the fact that there was an error in how the overall accuracy of the feedback was communicated to teachers during the initial presentation and in the tool's information screens might have influenced their perceptions of accuracy. We do not think this was a major concern because an analysis of the interviews indicated that teachers were not basing their perceptions of accuracy on the generic information provided, but rather by comparing their own judgments of each talk feature vs. the feedback provided. That judgement often referenced the lesson context, what the teacher believed to be their enduring practices/strengths, or simply their own recollection of the lesson (Roxanne: "…just from my memory of it, I want to say maybe it's something like 80% accurate.").

[13] We experimented with that in this study in the second interview, and users generally responded positively to specific examples.

Future work should build on the results of this feedback-response study, considering some of the design possibilities we have discussed. Future work should also investigate how to integrate the Teacher Talk Tool into existing professional development and teacher training efforts. Additionally, feedback for teacher learning is not the only potential use of this tool, and data from the tool could be aggregated to higher levels for other purposes (e.g., examining school-to-school variation in instruction; see e.g., Kelly et al., 2020a, b).

Thinking about practical lessons learned in providing teachers with automated feedback to promote teacher learning, we would highlight two features. The first and most obvious lesson is to never include constructs with low face validity to participants. A construct like "instructional time" which might be very useful for research purposes (Kraft & Novicoff, 2024), but is less useful for individual feedback because it varies so little, should be avoided. In this case, because the feedback was also norm-referenced, and participants received "low" scores even as that was not true in any meaningful sense to them, the distracting effect of including this construct was exacerbated. While we can be much less certain, we believe a second lesson is that providing the precise, raw prevalence rates, without reference to or accompanied by norm-referenced criteria, is preferable to providing categorical scores like "low," "medium," and "high." One of the main reasons we chose to use categorical scores was because then the accuracy of automation could be conveyed to participants in a simple metric. But that now strikes us as insufficient and poor justification for a design feature that made the feedback seem more evaluative than it needed to be or we intended it to be. The tradeoffs involved therein should be investigated in future research.

In conclusion, this study was one of the first to provide teachers with automated feedback on instructional practice based on a fine-grained, agnostic coding of teacher talk. As an external source of information, we found that teachers found the feedback to be accurate and this feedback prompted a high degree of reflection, focusing teachers' attention on the measured talk constructs. This feedback also led teachers to engage in a process of sense-making, linking the measured talk features to teacher-, student-, and lesson-level variables. To a lesser extent, we also found evidence of feedback uptake, including goal-setting, while evidence of strategy use was limited.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40593-024-00417-x.

Author Contributions Author 1: Conceptualization; Data Collection; Methodology-Surveys; Analysis; Writing-Composing, Review, and Editing. Author 2: Analysis; Editing. Author 3: Methodology-Software and Automation; Data Collection-user support. Author 4: Conceptualization; Methodology-Software and Automation; Writing-Composing, Review, and Editing.

Funding This research was supported by the National Science Foundation (NSF IIS 1735785). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the funding agencies.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Ahuja, K., Kim, D., Xhakaj, F., Varga, V., Xie, A., Zhang, S., Townsend, J. E., Harrison, C., Ogan, A., & Agarwal, Y. (2019). EduSense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3*(3), 1–26.

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse, arXiv preprint arXiv:2208.04715.

Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations*. John Wiley & Sons.

Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., ... & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics*, *12*, 423–436.

Aucejo, E., Coate, P., Fruehwirth, J. C., Kelly, S., & Mozenter, Z. (2022). Teacher effectiveness and classroom composition: Understanding match effects in the classroom. *The Economic Journal, 132*, 3047–3064.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 50–97). Jossey-Bass.

Blanchard, N., Brady, M., Olney, A., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S., & D'Mello, S. K. (2015). A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Proceedings of the 17th international conference on artificial intelligence in education* (AIED 2015) (pp. 23–33). Springer-Verlag.

Brett, J. F., & Atwater, L. E. (2001). 360° feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology, 86*, 930–942.

Camburn, E. M. (2010). Embedded teacher learning opportunities as a site for reflective practice: An exploratory study. *American Journal of Education, 116*, 463–489.

Camburn, E. M., & Han, S. W. (2015). Infrastructure for teacher reflection and instructional change: An exploratory study. *Journal of Educational Change, 16*, 511–533.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*, 1233–1267.

Cao, J., Ganesh, A., Cai, J., Southwell, R., Perkoff, M., Regan, M., Kann, K., Martin, J., Palmer, M., & D'Mello, S. K. (2023). A comparative analysis of automatic speech recognition errors in small group classroom discourse. In Proceedings of the ACM International Conference on User Modeling, Adaptation and Personalization (UMAP 2023) (pp. 250–262). ACM.

Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S., & Fine, J. G. (2013). English teacher candidates developing dialogically organized instructional practices. *Research in the Teaching of English, 47*, 212–246.

Chawla, N., Gabriel, A. S., da Motta Veiga, S. P., & JSlaughter, J. E. (2019). Does feedback matter for job search self-regulation? It depends on feedback quality. *Personnel Psychology, 72*, 513–541.

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*, 642–680.

Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter?* (REL 2017–190). Regional Educational Laboratory Central.

Chiu, J. L., Bywater, J. P., & Lilly, S. (2022). The role of AI to support teacher learning and practice: A review and future directions. In F. Ouyang, P. Jiao, B. McLaren, & A. Alavi (Eds.), *Artificial intelligence in STEM education: The paradigmatic shifts in research, education, and technology* (pp. 163–173). CRC Press.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*, 947–967.

Close, K., Amrein-Beardsley, A., & Collins, C. (2018). *State-level assessments and teacher evaluation systems after the passage of the every student succeeds act: Some steps in the right direction*. National Education Policy Center.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*, 378–387.

Colestock, A., & Sherin, M. G. (2009). Teachers' sense-making strategies while watching video of mathematics instruction. *Journal of Technology and Teacher Education, 17*, 7–29.

d'Anjou, B., Bakker, S., An, P., & Bekker, T. (2019). How peripheral data visualisation systems support secondary school teachers during VLE-supported lessons. In *Proceedings of the 2019 on designing interactive systems conference* (pp. 859–870).

D'Mello, S. K., Lehman, B., & Person, N. (2010). Expert tutors feedback is immediate, direct, and discriminating. In C. Murray & H. Guesgen (Eds.), *Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference* (pp. 595–560). AAAI Press.

D'Mello, S. K., Olney, A. M, Blanchard, N., Sun, X., Ward, B., Samei, B., & Kelly, S. (2015). Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. *Proceedings of the 17th ACM International Conference on Multimodal Interaction* (ICMI 2015) (Multimodal Learning Analytics Grand Challenge MLA'15). (pp. 557–566). ACM.

Dale, M., Godley, A., Capello, S., Donnelly, P., D'Mello, S., & Kelly, S. (2022). Toward the automated analysis of teacher talk in secondary ELA classrooms. *Teaching and Teacher Education, 110*, 103584.

Datta, D., Bywater, J. P., Phillips, M., Lilly, S., Chiu, J. L., Watson, G. S., & Brown, D. E. (2023). Classifying mathematics teacher questions to support mathematical discourse. In *International Conference on Artificial Intelligence in Education* (pp. 372–377). Springer Nature Switzerland.

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*. https://doi.org/10.3102/01623737231169270

Demszky, D. (2022). Using natural language processing to support student-centered education, Doctoral dissertation, Stanford University.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, arXiv:1810.04805.

Dwyer, C. A., & Stufflebeam, D. S. (1996). Evaluation for effective teaching. In D. Berliner & R. Calfee (Eds.), *Handbook of research in educational psychology.* Macmillan.

Ericsson, K. A., Krampe, R., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363.

Fadde, P. J., & Klein, G. A. (2010). Deliberate performance: Accelerating expertise in natural settings. *Performance Improvement, 49*(9), 5–14.

Gamoran, A., & Nystrand, M. (1992). Taking students seriously. In F. Newmann (Ed.), *Student engagement and achievement in American secondary schools.* Teachers College Press.

Gamoran, A., Nystrand, M., Berends, M., & Lepore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal, 32*, 687–715.

Gamoran, A., & Kelly, S. (2003) Tracking, instruction, and unequal literacy in secondary school English. In M. T. Hallinan, A. Gamoran, W. Kubitschek, and T. Loveless (Eds.), *Stability and Change in American Education: Structure, Processes and Outcomes* (pp. 109–126). Eliot Werner Publications.

Gerard, L., Wiley, K., Bradford, A., Chen, J. K., Lim-Breitbart, J., & Linn, M. (2020). Impact of a teacher action planner that captures student ideas on teacher customization decisions. In Proceedings of the 14th international society for learning sciences conference (pp. 2077–2084).

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(6), 1–32.

Grissom, J. A., Blissett, R. S., & Mitani, H. (2018). Evaluating school principals: Supervisor ratings of principal practice and principal job performance. *Educational Evaluation and Policy Analysis, 40*(3), 446–472.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 19*, 45–470.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29–48.

Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture, and Social Interaction, 25*, 100404.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. (Tech. Rep.). Bill & Melinda Gates Foundation, Measures of Effective Teaching Project.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*, 407–434.

Huang, G. Y., Chen, J., Liu, H., Fu, W., Ding, W., Tang, J., ... & Liu, Z. (2020). Neural multi-task learning for teacher question detection in online classrooms. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21* (pp. 269–281). Springer International Publishing.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*, 253–263.

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussion in mathematics classrooms: Using personalized automated feedback to support reflection and instructional change. *Teaching and Teacher Education, 112*, 103611.

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

Jensen, E., Pugh, S., & D'Mello, S. K. (2021). A deep transfer learning approach to automated teacher discourse feedback. In *Proceedings of the 11th Learning Analytics & Knowledge Conference (LAK 2021)*. ACM.

Kelly, S. (2007). Classroom discourse and the distribution of student engagement. *Social Psychology of Education, 10*, 331–352.

Kelly, S. (2008). Race, social class, and student engagement in middle school English classrooms. *Social Science Research, 37*, 434–448.

Kelly, S. (2023). Agnosticism in instructional observation systems. *Education Policy Analysis Archives, 31*(7). https://doi.org/10.14507/epaa.31.7493

Kelly, S., & Abruzzo, E. (2021). Using lesson-specific teacher reports of student engagement to investigate innovations in curriculum and instruction. *Educational Researcher, 50*, 306–314.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher, 47*, 451–464.

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. (2020a). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives, 28*, 62.

Kelly, S., Mozenter, Z., Aucejo, E., & Fruehwirth, J. (2020b). School-to-school differences in instructional practice: New descriptive evidence on opportunity to learn. *Teachers College Record, 122*(11), 1–38.

Klette, K., Blikstad-Balas, M., & Roe, A. (2017). Linking instruction and student achievement. *Acta Didactica, 11*(3), 10.

Korban, M., Youngs, P., & Acton, S. T. (2023). A Multi-Modal Transformer network for action detection. *Pattern Recognition, 142*, 109713.

Kraft, M. A., & Christian, A. (2019). *In search of high-quality evaluation feedback: An administrator training field experiment*. Ed-Working Paper 19–62, Annenberg Institute at Brown University, Providence, RI.

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*, 711–753.

Kraft, M. A., & Novicoff, S. (2024). Time in school: A conceptual framework, synthesis of the causal research, and empirical exploration. *American Educational Research Journal, 0*(0). https://doi.org/10.3102/00028312241251857

Langer, J. A. (2001). Beating the odds: Teaching middle and high school students to read and write well. *American Educational Research Journal, 38*, 837–880.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation, and Accountability, 31*, 61–95.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithmic appreciation: People prefer algorithmic to human judgement. *Organizational Behavior and Human Decision Processes, 151*, 90–103.

Lugini, L., Litman, D., Godley, A., & Olshefski, C. (2019). Annotating student talk in text-based classroom discussions, *arXiv preprint* arXiv:1909.03023.

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice, 34*(2), 34–46.

McKeown, M. G., & Beck, I. L. (2015). Effective classroom talk is reading comprehension instruction. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 51–62). American Educational Research Association.

Miles, Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis : a methods sourcebook*. (Fourth edition.). SAGE.

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' high-level comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*, 740–764.

Palonsky, S. B. (1986). *900 Shows a year: A look at teaching from a teacher's side of the desk*. McGraw-Hill.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

Pepper, M. J., Ehlert, M. W., Parsons, E. S., Stahlheber, S. W., & Burns, S. F. (2015). *Educator evaluations in Tennessee: Findings from the 2014 First to the Top survey*. Tennessee Consortium on Research, Evaluation, & Development. Vanderbilt.

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM Mathematics Education, 50*(3), 535–553.

Price, H. E. (2012). Principal-teacher interactions: How affective relationships shape principal and teacher attitudes. *Educational Administration Quarterly, 48*, 39–85.

Price, H. E. (2021). Weathering fluctuations in teacher commitment: Leaders relational failures, with improvement prospects. *Journal of Educational Administration, 59*, 493–513.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*, 4–15.

Quintelier, A., De Maeyer, S., & Vanhoof, J. (2020). Determinants of teachers' feedback acceptance during a school inspection visit. *School Effectiveness and School Improvement, 31*, 529–547.

Resnick, L. B., Asterhan, C. S. C., Clarke, S. N., & Schantz, F. (2018). Next generation research in dialogic learning. In G. E. Hall, L. F. Quinn, & D. M. Gollnick (Eds.), *Wiley handbook of teaching and learning* (pp. 323–338). Wiley-Blackwell.

Reznitskaya, A., Anderson, R. C., McNurlen, B., Nguyen-Jahiel, K., Archodidou, A., & Kim, S.-O. (2001). Influence of oral discussion on written argument. *Discourse Processes, 32*, 155–175.

Sankaranarayanan, S., Kandimalla, S. R., Hasan, S., An, H., Bogart, C., Murray, R. C., ... & Rosé, C. (2020). Agent-in-the-loop: conversational agent support in service of reflection for learning during collaborative programming. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21* (pp. 273–278). Springer International Publishing.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*, 377–400.

Shernoff, D. J. (2013). *Optimal learning environments to promote student engagement*. Springer.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Song, Y., Lei, S., Hao, T., Lan, Z., & Ding, Y. (2021). Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research, 59*, 496–521.

Southwell, R., Pugh, S., Perkoff, E. M., Clevenger, C., Bush, J. B., Lieber, R., ... & D'Mello, S. (2022). Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms. International Educational Data Mining Society.

Stigler, J. W., & Miller, K. F. (2018). Expertise and expert performance in teaching. In *The Cambridge handbook of expertise and expert performance* (pp. 431–452). Cambridge University Press. https://doi.org/10.1017/9781316480748.024

Suresh, A., Sumner, T., Huang, I., Jacobs, J., Foland, B., & Ward, W. (2018). Using deep learning to automatically detect talk moves in teachers' mathematics lessons. In *2018 IEEE International Conference on Big Data (Big Data)*, 5445–5447.

Suresh, A., Sumner, T., Jacobs, J., Foland, B., & Ward, W. (2019). Automating analysis and feedback to improve mathematics teachers' classroom discourse. Proceedings of the AAAI Conference on Artificial Intelligence.

Suresh, A., Jacobs, J., Perkoff, M., Martin, J., & Sumner, T. (2022). Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. 17th Workshop on Innovative Use of NLP for Building Educational Applications.

Taylor, B. M., Pearson, P. D., Peterson, D. P., & Rodriguez, M. C. (2005). The CIERA School change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly, 40*, 40–69.

Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L. C. (2023). Utilizing natural language processing for automated assessment of classroom discussion. In International Conference on Artificial Intelligence in Education (pp. 490–496). Springer Nature Switzerland.

Tschannen-Moran, M., & Hoy, W. (1998). Trust in schools: A conceptual and empirical analysis. *Journal of Educational Administration, 36*, 334–352.

van de Grift, W. J. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*, 295–311.

Van Maele, D., & Van Houtte, M. (2009). Faculty trust and organizational school characteristics: An exploration across secondary schools in Flanders. *Educational Administration Quarterly, 45*, 556–589.

Vanover, C., Mihas, P., & Saldaña, J. (Eds.). (2021). *Analyzing and interpreting qualitative research: After the interview*. SAGE Publications.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems, 30: Annual conference on neural information processing systems 2017* (pp. 5998–6008).

White, M. C. (2018). Rater performance standards for classroom observation measures. *Educational Researcher, 47*, 492–501.

White, M., & Klette, K. (2023). What's in a score? Problematizing interpretations of observation scores. *Studies in Educational Evaluation, 77*, 101238.

Wieczorek, D., Aguilar, I., & Mette, I. (2022). System-level leaders' local control of teacher supervision and evaluation under every student succeeds act. *AASA Journal of Scholarship & Practice*, *19*(3), 10–31.

Wilkinson, I. A., Soter, A., & Murphy, P. (2010). Developing a model of quality talk about literary text. In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 142–169). Guilford.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). HuggingFace's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771

Wylie, E. C., & Lyon, C. J. (2020). Developing a formative assessment protocol to support professional growth. *Educational Assessment, 25*(4), 314–330.