Research Handbook on Classroom Observation

Edited by

Sean Kelly

Professor, Department of Educational Foundations, Organizations, and Policy, University of Pittsburgh, USA

ELGAR HANDBOOKS IN EDUCATION

EE Edward Elgar
PUBLISHING

Cheltenham, UK • Northampton, MA, USA

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

Authorised representative in the EU for GPSR queries only: Easy Access System Europe – Mustamäe tee 50, 10621 Tallinn, Estonia, gpsr.requests@easproject.com

A catalogue record for this book
is available from the British Library

Library of Congress Control Number: 2024952811

This book is available electronically in the **Elgar**online
Sociology, Social Policy and Education subject collection
https://doi.org/10.4337/9781035321544

ISBN 978 1 0353 2153 7 (cased)
ISBN 978 1 0353 2154 4 (eBook)

20. Automating feedback from recorded instructional observations: using AI to detect and support dialogic teaching

Jennifer Jacobs, Abhijit Suresh, Brandon M. Booth, Tamara Sumner, Jeffrey Bush, Chelsea Brown and Sidney K. D'Mello

INTRODUCTION

The Importance of Classroom Discourse

A growing number of proponents are calling for classrooms that encourage student-driven academic discourse, called dialogic classrooms. Although such classrooms are relatively rare (Applebee et al., 2003), there is a rising interest in delineating, encouraging, and studying them. Despite the lack of large-scale studies, mounting evidence suggests dialogic classrooms can support student success in the form of increased agency, engagement, motivation, and learning gains (Böheim et al., 2021; Howe et al., 2019; Kelly & Abruzzo, 2021; Resnick et al., 2015; Webb et al., 2019). Dialogic classrooms are generally consistent with the long-standing recommendations of educators across disciplines, such as the Principles and Standards for School Mathematics (National Council of Teachers of Mathematics, 2000), the Framework for K–12 Science Education (National Research Council, 2012), and the Common Core State Standards for English Language Arts (NGA & CCSSO, 2010). These recommendations include positioning students to actively communicate their thinking to others in the classroom, engage in oral and written argumentation, and participate in extended collaborative discussions.

Although there is no single, agreed-upon definition of dialogic teaching (Alexander, 2018), most descriptions are framed around two broad aspects of classroom activities: (1) the distribution of talk between teachers and students and (2) the nature of the ideas put forward representing multiple viewpoints (Hofmann & Ruthven, 2018). Resnick and colleagues offer a succinct yet broadly encompassing description of this pedagogical approach that is widely referenced in the literature and forms the foundation for accountable talk, which grounds the AI application discussed in this chapter. Resnick and colleagues posit:

> In dialogic learning, students think out loud about a complex problem that requires collaboration: noticing something about the problem, questioning a surprising finding, or articulating, explaining, and reflecting upon their own reasoning…. Overall, the teacher's goal is to sustain a teacher-led but student-owned process of shared reasoning that ultimately results in a more fully developed, evidence-backed conclusion, solution, or explanation. (Resnick et al., 2018, pp. 325–326)

Translating the philosophy behind dialogic learning into everyday classroom instruction represents a significant shift for most teachers (Herbel-Eisenmann et al., 2013; Lyle, 2008). Transitioning from leading more authoritative and monologic lessons to facilitating dialogic lessons requires significant effort, intentionality, and skill (Hennessy & Davies, 2019; Webb et al., 2014). Research suggests that dialogic teaching requires teachers to adopt a purposeful mindset and a new behavioral repertoire, taking on a dialogical stance that includes a genuine commitment to soliciting students' ideas, encouraging the exchange of ideas, and promoting student agency (Davies et al., 2017). Unfortunately, much of the schooling infrastructure works against dialogic teaching, such as large class sizes, curriculum selection, assessment practices, and other institutional constraints (Lefstein & Snell, 2013). At the same time, there is an expanding literature on how teachers can move towards the purposeful incorporation of more dialogic practices (Walshaw & Anthony, 2008).

Opening a dialogic space for learners can take the form of using intentionally designed tasks, activities, and classroom routines to foster dialogue, as well as asking questions and providing other linguistic prompts (Wegerif, 2013). In this chapter, we focus on the discursive moves that invite student contributions and encourage the collaborative construction of knowledge. These moves encourage dialogic learning by inviting multiple voices to enter and shape the knowledge-building process during periods of joint work on a given topic (Mercer et al., 2019). Multiple codifications of dialogic teaching moves have been proposed, with accountable talk theory being the most widely used at present (Tao & Chen, 2023). Accountable talk theory has helped to articulate and define several talk moves that can promote active engagement in rich discussions (O'Connor et al., 2015; Resnick et al., 2018).

Accountable talk supports the educative potential of classroom discourse (Teo, 2019) by highlighting three dimensions of accountability in instructional practice: accountability to the learning community, accountability to content knowledge, and accountability to rigorous thinking. At the heart of accountable talk is the notion that teachers should organize inclusive and equitable discussions in a rigorous learning environment (Michaels et al., 2010; O'Connor & Michaels, 2019). Talk moves are a tool to generate such conversations by encouraging students to contribute and listen to each other, engage with the math content, and dig deeply into their own reasoning. For example, teachers can use moves such as *pressing for reasoning* or *revoicing* a student's idea to foster accountability to rigorous thinking. Correspondingly, students can use moves such as *providing an explanation or reasoning* or *linking* their ideas to those of their peers (Candela et al., 2020). An increasingly common way to direct teachers' and students' attention to talk moves is by prominently displaying accountable talk stems, sentence starters, and similar resources in classrooms (e.g., Wagganer, 2015; Walter, 2018).

Using dialogic talk moves helps to enable an instructional shift from teacher-directed recitation to "true discussions" in which, in their fullest expression, knowledge is shared, negotiated, and constructed (Cazden, 2003). Moreover, by scaffolding conversations in which students play a central and purposeful role, teachers help to socialize children into a particular academic enterprise in which they are legitimate and essential participants (O'Connor & Michaels, 1996). Researchers have raised concerns regarding inequities in classroom discourse, including differences in engagement corresponding to academic achievement status (Kelly, 2008) as well as disparities in the quality of talk and opportunities to participate by gender, race, ethnicity, and other social markers (Reinholz & Shah, 2018). Establishing learning environments conducive to elaborated and sophisticated student talk is particularly important for emerging multilingual students. Such environments foster language development and promote attention to and development of the resources (e.g., gestures, objects, everyday experiences) multilingual students use to communicate in classroom discussions (Moschkovich, 2002).

Approaches to Observing Classroom Discourse

It is possible to observe the discourse in classroom lessons with various goals and applications in mind, including determining the degree to which the lessons are dialogic. As Lefstein and Snell (2013) argue, “We can ask of any discourse event who speaks to whom, about what, how often and for how long, and on the basis of the answers to those questions make judgments about how dialogic the event is and in what ways” (p. 15). Much like classroom observations in general, observations of the discourse can help to identify patterns and variations, document change over time, define the nature of “effective” discourse, and provide feedback that might inform professional learning and data-driven decision making (Calcagni et al., 2023).

Capturing conversational interactions through video and audio has long been critical for all types of analyses attending to collaborative dialogue (Erickson, 2011). Rapid advances in recording technologies enable teachers to easily and unobtrusively self-record their classrooms, with examples such as “classroom robotics” designed specifically for educators to film their instructional environments (Franklin et al., 2018). In addition, expanded storage capacities and newly developed online repositories aid in managing, sharing, analyzing, and learning from classroom recordings, affording a plethora of educative uses (Ramos et al., 2022).

A wide variety of approaches, spanning the spectrum from qualitative and quantitative methodologies, have been applied to the analysis of video- or audio-recorded classroom discourse data. Traditionally, fine-grained analyses of classroom discourse have relied primarily on detailed human annotation. Micro-level, turn-by-turn coding is commonly used in qualitative analyses, which enables careful attention to detailed information, such as how the co-construction of knowledge occurs in a lesson (Hennessy et al., 2023). However, such efforts are human resource-intensive and difficult to scale to large quantities of data and, as such, are mostly confined to research projects. Approaches that support scaling these analyses to larger datasets could drive new insights, for instance, by connecting assessments of classroom dialogue to other variables of interest, such as measures of student performance (Howe & Abedin, 2013).

A significant barrier to using recorded lessons for automated analyses and corresponding teacher feedback has been the difficulty of recording sufficiently high-quality audio tracks. High-quality audio input is critical in order to generate accurate analytics about the specific nature of teachers’ and students’ utterances during a lesson. School classrooms are noisy environments, making it difficult to capture speech and language components of teacher and student interactions with high fidelity and without posing a major disruption to teachers and students (D’Mello et al., 2015). Building on advances in recording technologies, recent studies have demonstrated that classroom audio of sufficient quality can be collected with minimal imposition (Bokhove & Downey, 2018; Donnelly et al., 2016; Jensen et al., 2020). Being able to reliably and robustly record teacher and student classroom dialogue is a critical advance, enabling new computational methods supporting the automated analysis of classroom recordings.

In the past few years, there has been an explosion in efforts to develop advanced algorithms in natural language processing (NLP) in the form of AI language models such as ChatGPT, exemplifying the capability of these models to perform complex tasks with high accuracy. Building on these advances in the converging areas of automatic speech recognition, natural language processing, and machine learning, recent research has shown that the development and training of large language models to automate and scale discourse analyses is feasible (Song et al., 2021; Suresh et al., 2021). Working from recordings of speech from

K–12 classroom environments, researchers have developed a variety of AI models to reliably detect discursive features such as accountable talk, instructional talk, authentic teacher questions, elaborated evaluation, talk moves, and uptake (Datta et al., 2023; Demszky et al., 2023a; Donnelly et al., 2017; Jensen, E. et al., 2021; Kelly et al., 2018; Suresh et al., 2021; Tran et al., 2023). However, to date, only a few of these models have been integrated into deployed, teacher-facing professional learning tools.

The Potential of Automated Discourse Analyses to Support Teacher Learning

It is widely recognized that receiving personalized feedback on one's work performance has a powerful influence on behavior and helps drive continuous improvement (Bryk et al., 2015; Hattie & Timperley, 2007). An extensive body of research suggests that formative feedback, or information that is intended to promote learning, should be non-evaluative, supportive, timely, and specific (Shute, 2008). The current gold standard for providing personalized feedback to teachers on their classroom instruction involves expert human observers, an effort that tends to be time-consuming, expensive, and challenging to provide at scale (van der Lans et al., 2016). In addition, human observers can be highly subjective, imprecise, and inconsistent, even when using standardized observation protocols (Kelly et al., 2020). Automated approaches to classifying instructional practice have fewer human biases and can be implemented with lower costs and higher turnaround speeds, offering a potentially transformative approach to classroom observation (Liu & Cohen, 2021).

Recent research highlights the promising role that automated, data-driven feedback based on classroom recordings can play in supporting teacher learning about dialogic instruction. For example, Demszky and colleagues (2023a) created a tool called M-Powering Teachers that uses NLP models to provide online computer science college instructors with autogenerated information about their uptake of students' contributions. Demszky et al. (2023a) describe uptake as a "high leverage dialogic teaching practice" (p. 1) that includes revoicing, question answering, elaboration, and other strategies indicating teachers are using students' ideas as resources. The research team conducted a randomized controlled trial and found that receiving this feedback increased instructors' uptake of contributions by 13% and improved students' satisfaction with the course. Similarly, positive results were detected when the tool was used by research mentors who received feedback as they worked with high school students through an online platform (Demszky & Liu, 2023). Another example is the Teacher Talk Tool developed by Kelly and colleagues (under review) that provides automatic feedback on aspects of teacher discourse in English Language Arts classes. A small-scale longitudinal feedback-response study with five teachers found that the Teacher Talk Tool promoted teacher reflection and focused attention on high-leverage discourse moves, though evidence of uptake was not investigated.

The authors of this chapter developed an AI-based tool that provides mathematics teachers with automated, personalized feedback on their classroom discourse in alignment with accountable talk theory. This tool, called TalkMoves, generates information about the specific discourse moves used by both teachers and students during recorded lessons. A longitudinal pilot of the TalkMoves application points to its utility value for teachers, including a positive impact on their discourse practices over time (Jacobs et al., 2022, 2024; Suresh et al., 2024). Similar to the tools developed by Demszky and Kelly's teams, the TalkMoves application is fully automated and supports learning about dialogic practices through a minimal intervention approach.

In the next two sections of this chapter, we describe the TalkMoves application in detail, including the training data it is built from, the application architecture, NLP model performance, and the feedback dashboard. We present findings from a field study of the TalkMoves application, including classroom teachers' documented use and perceived efficacy of the application and observed changes in their lessons. We then discuss how the TalkMoves application was revised and updated for a different instructional context involving novice tutors and their instructional coaches.

USE CASE 1: THE TALKMOVES APPLICATION FOR TEACHERS

Application Overview

The TalkMoves application is a deployed web-based system that uses automatic speech recognition and deep learning models to analyze classroom recordings and detect the presence of teacher and student talk moves, drawing on accountable talk theory (Jacobs et al., 2022; Suresh et al., 2024). The system architecture involves an end-to-end, fully automated design to capture and process discursive information from recorded audio. This infrastructure includes data management and storage, a processing pipeline, and feedback generation. The application consists of three interrelated components: (1) a cloud-based big data infrastructure to manage and process lesson recordings, (2) automated speech recognition and NLP models to classify talk moves, and (3) a personalized dashboard to visually display each teacher's feedback analytics for their individual lessons and their lessons over time.

First, teachers generate and upload recordings of lessons directly into the TalkMoves web portal or through classroom-assisted technologies such as the Swivl (Franklin et al., 2018). The system asynchronously processes each recording through a sequence of steps, starting with converting the audio into a written transcript using automatic speech recognition services. Recordings with high-quality audio inputs are critical for this speech processing to produce accurate transcripts that form the basis of all downstream analyses (Jensen et al., 2020). The transcripts are then processed in a variety of ways, for example, to include timestamps and to indicate whether utterances are produced by the teacher or a student. Deep learning AI models analyze the transcripts to determine which sentence corresponds to a teacher- or student-generated accountable talk move.

Along with the resulting predictions from the AI models, additional analytics are applied to calculate other discursive features, such as how much talk came from the teacher versus the students, the degree to which the teacher incorporated wait time in their discourse, and students' use of math content words. Finally, the system compiles all this feedback and visually displays it on a personalized dashboard. Shortly after uploading a recording,[1] teachers receive an email from the application notifying them that their lesson has been processed and providing a direct link to that lesson's feedback interface. To ensure the privacy of both the teachers and their students, the application is password-protected and structured such that teachers can only view the feedback on their own lessons.

The AI models used within the TalkMoves application were custom-built by the research team utilizing a carefully crafted, human-labeled dataset to train a model with high accuracy for predicting specific accountable talk moves (Suresh et al., 2022a). This mode of training from labeled datasets is referred to as supervised learning in the realm of AI and machine learning. The training data were sourced from real-world, human-generated classroom transcripts based on recordings of K–12 mathematics classrooms. All the transcripts were annotated with six teacher talk moves (*keeping everyone together*, *getting students to relate to another's ideas*, *restating*, *pressing for accuracy*, *revoicing*, and *pressing for reasoning*) and four student talk moves (*relating to another student*, *asking for more information*, *making a claim*, and *providing reasoning*), as shown in Tables 20.1a and 20.1b. These talk moves were selected for inclusion based on suggestions from accountable talk experts and the degree to which they lent themselves to the construction of accurate computational models (i.e., they were relatively frequent, and trained humans could establish high reliability). The selected talk moves represent the three categories of accountable talk: accountability to the learning community, content knowledge, and rigorous thinking.

For each teacher and student sentence in the transcript, the NLP deep learning models classify teacher and student sentences into discourse classes (or labels), aligning with the 10 target talk moves and distinguishing between sentences with and without talk moves. The model used in the deployed application was selected based on a series of exhaustive experiments with various state-of-the-art model architectures for sentence-level classification, including transformers such as BERT (Devlin et al., 2019), Roberta (Liu et al., 2019), and BigBird (Zaheer et al., 2020). These large language models, also known as transformers, have exploded in popularity in the past decade because they can capture and apply context information across a wide range of data. Experiments with different transformer architectures yielded computational models with classification performance generally on par with well-trained humans, demonstrating the reliability and robustness of artificial intelligence algorithms applied to a

Table 20.1a Teacher talk moves included in the TalkMoves application

Category	Talk move	Description	Example
Learning Community	Keeping everyone together	Prompting students to be active listeners and orienting students to each other	"What did Eliza just say her equation was?"
Learning Community	Getting students to relate to another's ideas	Prompting students to react to what a classmate said	"Do you agree with Juan that the answer is 7/10?"
Learning Community	Restating	Repeating all or part of what a student said word for word	"Add two here."
Content Knowledge	Pressing for accuracy	Prompting students to make a mathematical contribution or use mathematical language	"Can you give an example of an ordered pair?"
Rigorous Thinking	Revoicing	Repeating what a student said but adding on or changing the wording	"Julia told us she would add two here."
Rigorous Thinking	Pressing for reasoning	Prompting students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations	"Why could I argue that the slope should be increasing?"

Table 20.1b Student talk moves included in the TalkMoves application

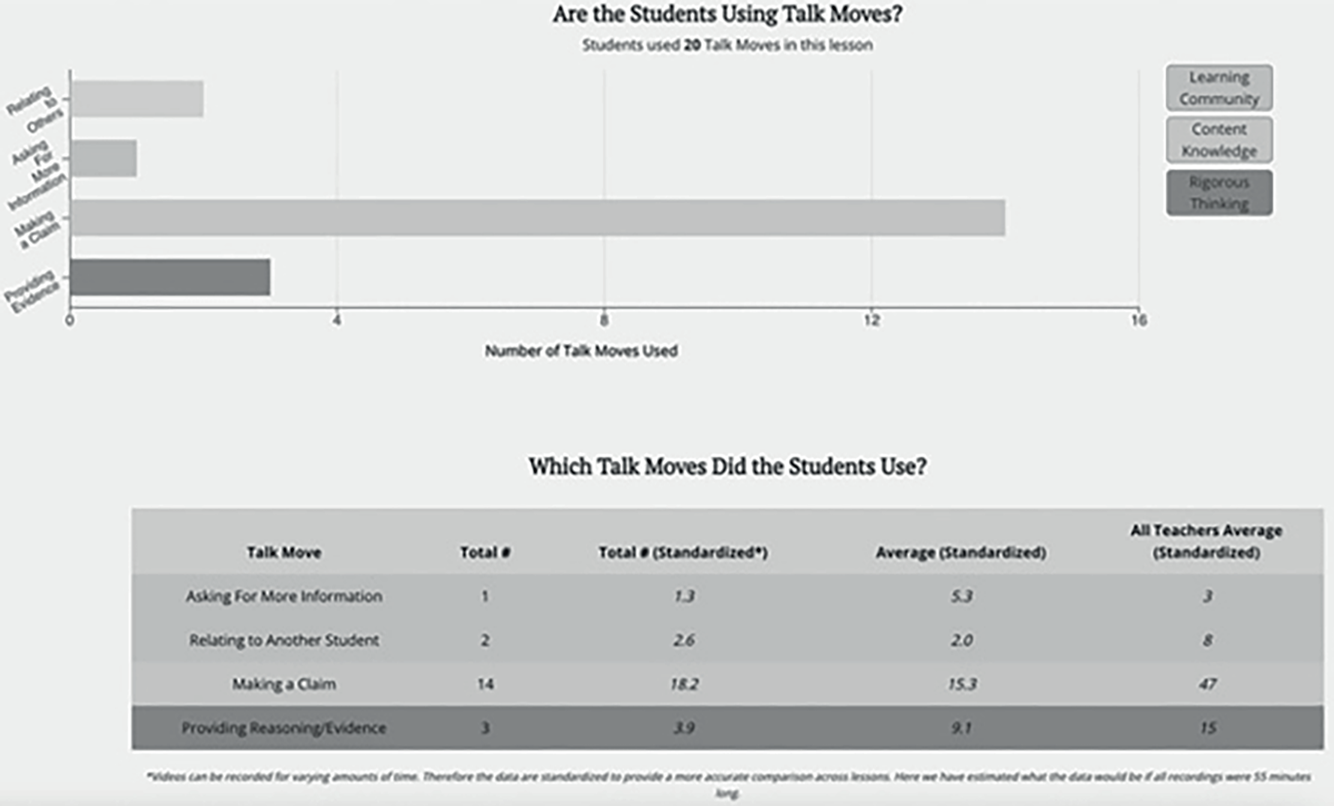
Category	Talk move	Description	Example
Learning Community	Relating to another student	Using, commenting on, or asking questions about a classmate's ideas	"I didn't get the same answer as her."
Learning Community	Asking for more info	Student requests more info, says they are confused or need help	"I don't understand number four."
Content Knowledge	Making a claim	Student makes a math claim, factual statement, or lists a step in their answer	"X is the number of cars."
Rigorous Thinking	Providing evidence or reasoning	Student explains their thinking, provides evidence, or talks about their reasoning	"You can't subtract 7 because then you would only get 28 and you need 29."

challenging educational context (see Suresh et al., 2019, 2021 for details). Of note is that the models using the TalkMoves application, as well as the corresponding training data, have been publicly released (Suresh et al., 2022a) and several other research teams have further experimented with and applied them to different instructional contexts (Balyan et al., 2022; Booth et al., 2024; Wang et al., 2023).

The nature and presentation of the TalkMoves feedback was generated through a collaborative design (co-design) process (Penuel et al., 2007) undertaken by a group of teachers, mathematics educators, learning scientists, and computer scientists. The resulting interfaces convey information using intuitive graphs and charts, highlighting patterns within individual lessons and longitudinally across multiple lessons. See Figure 20.1 for two example data displays of student talk moves. These displays show how often each type of student talk move occurred (displayed here in grayscale; teacher participants viewed colors indicating the accountable talk category) in the target lesson, along with the average frequency of these talk moves within all of the individual teacher's lessons and across the full set of teachers who used the application. The averages provide users with points of comparison but do not offer a "target" or make a judgment about the frequency of the talk moves within their lessons.

TalkMoves Pilot

We conducted a small, longitudinal pilot study of the TalkMoves application to explore teachers' use of the tool, their perceptions of its utility, and its impact on their classroom instruction. Twenty-one teachers from two school districts in the western United States volunteered to participate in the pilot beginning in Fall 2019. The teachers spanned grades 4–12, with most teaching upper elementary school (71%). The participants varied in their amount of classroom teaching experience (ranging from 4–32 years), but on average, were a relatively experienced group (M=15). Twelve teachers continued participating in the pilot for a second school year (2020–21). Like the full group, most of these continuing teachers taught elementary school (67%), and their average teaching experience was 16 years.



Are the Students Using Talk Moves?

Students used **20** Talk Moves in this lesson

Which Talk Moves Did the Students Use?

Talk Move	Total #	Total # (Standardized*)	Average (Standardized)	All Teachers Average (Standardized)
Asking For More Information	1	*1.3*	*5.3*	*3*
Relating to Another Student	2	*2.6*	*2.0*	*8*
Making a Claim	14	*18.2*	*15.3*	*47*
Providing Reasoning/Evidence	3	*3.9*	*9.1*	*15*

**Videos can be recorded for varying amounts of time. Therefore the data are standardized to provide a more accurate comparison across lessons. Here we have estimated what the data would be if all recordings were 55 minutes long.*

Figure 20.1 *Interfaces displaying feedback about the occurrence of student talk moves during a selected lesson (displayed here in grayscale)*

During the first school year of the study, the teachers self-recorded and uploaded a total of 210 mathematics lessons (M=10 lessons per teacher). The COVID-19 pandemic disrupted the school year and likely was part of the reason why only about half of the teachers continued to use the application for a second year. Nevertheless, we were pleased that 12 teachers continued to participate and recorded a total of 163 mathematics lessons (M=14 lessons per teacher), some of which were held remotely on a video conferencing platform. Teachers recorded their in-person lessons using a Swivl device provided by the research team; they recorded online lessons using Zoom or Google Meet. The TalkMoves application processed each uploaded lesson and generated emails notifying teachers that feedback was available to view on their personalized dashboards.

As shown in Table 20.2, teachers perceived the TalkMoves application as useful and informative for their classroom practice in both years of the study. In Year 1, most teachers felt that the data tool was at least "somewhat" useful (86%) and that the feedback prompted them to change their instruction at least "to some degree" (90%). Perhaps not surprisingly, the teachers who elected to continue using the application for a second year had slightly more positive initial perceptions than the full sample. In Year 2, almost all of the teachers felt that the tool was "relevant" or "very relevant" to their everyday teaching (92%), and all expressed that the feedback informed their practice "somewhat" or "a lot" (100%).

The pilot study also explored changes in classroom discourse based on lesson recordings from the teachers who participated over two school years. These 12 teachers showed notable increases in both their own talk moves and their students' talk moves. Tables 20.3a and 20.3b show the results from the Wilcoxon signed ranks test for nonparametric data comparing the average frequency of talk moves from Fall 2019 to Spring 2021. There was a significant ($p < 0.05$) or nearly significant ($p < 0.06$) increase in three of the six teacher talk moves and all four of the student talk moves. Due to the small sample size and the lack of a comparison group, these analyses should be considered exploratory, and their generalizability is limited. However, they suggest that the teachers were motivated to engage with the application and use their data as a catalyst for self-reflection, eventually leading to observable changes in their everyday practice.

Table 20.2 Teachers' perceptions of the application's usefulness and how much it informed their practice in both years of the study

Teachers' Perceptions	**Y1 All Teachers (n=21) mean (SD)**	**Y1 Continuing Teachers (n=12) mean (SD)**	**Y2 Continuing Teachers (n=12) mean (SD)**
Usefulness of the feedback*	3.2 (0.8)	3.4 (0.7)	4.0 (1.2)
Informed practice**	3.6 (1.2)	3.9 (0.8)	3.9 (0.7)

Note: *Teachers' survey responses about how useful/relevant the feedback they received was (1 = not at all useful/relevant, 5 = very useful/relevant).
**Teachers' survey responses about how much the feedback informed their practice (1 = not at all, 5 = a lot).

Table 20.3a Change in the average frequency of teacher talk moves per lesson from Fall 2019 to Spring 2021

Talk Move Label	Fall 2019 Mean Frequency	Spring 2021 Mean Frequency
All teacher talk moves	100	129**
Keep students together	53	64+
Get students to relate	2	4
Restating	2	3
Press for accuracy	37	44+
Revoicing	4	11***
Press for reasoning	2	2

Note: [a]Averages were calculated by first determining the average for each teacher and then averaging across teachers. All data were normalized for a 55-minute lesson.
+p<.06, *p<.05, **p<.01, ***p<.001.

Table 20.3b Change in the average frequency of student talk moves per lesson from Fall 2019 to Spring 2021

Talk Move Label	Fall 2019 Mean Frequency	Spring 2021 Mean Frequency
All student talk moves	27	51*
Relate to another student	2	5+
Ask for info	2	4*
Make a claim	14	25*
Provide evidence	8	17*

Note: +p<.06, *p<.05.

Use of the Application for Individual Reflection

In this pilot of the TalkMoves application, teachers reviewed their classroom discourse data as part of an independent, cyclical process. The application provided an opportunity for individual reflection and learning for self-motivated teachers who recorded lessons and then looked over their feedback. The information provided to teachers in the form of visual analytics was intended to serve as a mirror rather than as an evaluation or outsider's call to action. At the same time, sense-making and generating ideas for instructional change were entirely placed on individual users.

As depicted in Figure 20.2 (left), the ideal use of the TalkMoves application during the pilot study was a cycle that involved teachers (1) reviewing and making sense of the feedback provided by the application, (2) reflecting on appropriate instructional changes based on the feedback, and (3) purposefully implementing instructional changes. Each newly uploaded lesson represented an opportunity for a teacher to review additional data and consider their progress

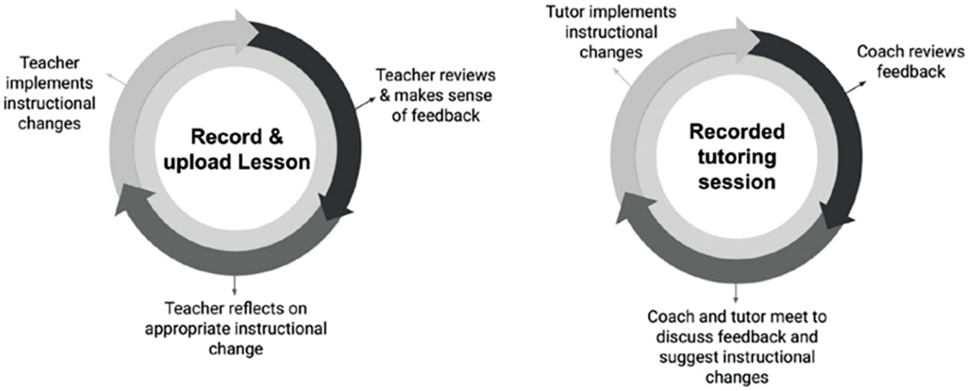

Teacher implements instructional changes
Record & upload Lesson
Teacher reviews & makes sense of feedback
Teacher reflects on appropriate instructional change
Tutor implements instructional changes
Recorded tutoring session
Coach reviews feedback
Coach and tutor meet to discuss feedback and suggest instructional changes

Figure 20.2 Intended use of the TalkMoves application by teachers working independently (left), current use of the application by tutors working with coaches (right)

on the metrics provided by the application. Certainly, the application has the potential to serve as a vehicle for shared inquiry by teachers, for instance, by structuring collective reviews of the feedback during professional learning workshops or instructional coaching sessions. In the next section, we will discuss how a revised version of the TalkMoves application was used by tutors in close consultation with instructional coaches.

USE CASE 2: THE TALKMOVES APPLICATION ADAPTED FOR TUTORS

The Tutoring Context

Building on the promise of the TalkMoves application for teachers, the tool was adapted for use by mathematics tutors. Human tutoring has consistently been found to yield impressive benefits for students. According to a meta-analysis of 96 randomized evaluations of tutoring programs, the vast majority of studies reported statistically significant impacts on student learning, with an impressive overall pooled effect size of 0.37 SD (Nickow et al., 2020). In the past few years, tutoring has become more widely accessible, particularly as a means of providing underserved youth with supplementary instruction as part of their regular school activities. A recent survey from a national sample of US public schools found that 59% provide standard tutoring (less than three times per week) and 37% provide high-dosage tutoring (three or more times per week) (Institute of Education Sciences, 2022).

The adapted tool leverages the AI models in the original application, fine-tuning them on a new dataset and customizing the application interfaces for use in tutoring contexts. This effort is part of a partnership with Saga Education, a large non-profit provider of mathematics tutoring services to high schoolers in Title 1 schools (i.e., public schools with predominantly low-income and historically marginalized student populations). Saga's program includes a partnership with AmeriCorps, such that individuals can complete their year of service as full-time tutors. Most tutors employed by Saga are "paraprofessionals"; they have college degrees but lack formal training in education and are not licensed or certified to be teachers. The tutors primarily work for one school year, during which they receive ongoing, individualized support and professional development from an instructional coach.

Participating schools build a tutoring period into their students' course schedules as part of Saga's high-dosage tutoring model, and students work with a human tutor every other day. Although the students are physically present in their school classroom, many tutors work remotely and interact with their assigned students via Saga's virtual tutoring virtual workspace, called Saga Connect. This shared workspace integrates video conferencing with digital whiteboards, calculators, and other tools to enable detailed mathematical representations and collaborative problem-solving. Tutors work with groups of 2–4 students at a time, providing the opportunity for collaborative and mathematically rich discourse.

Multiple evaluation studies have documented the efficacy of Saga's program for significantly improving students' learning in math and their persistence in school (Guryan et al., 2021; Nickow et al., 2020). School-based high-dosage tutoring is rapidly growing in popularity, and the paraprofessional tutoring workforce is scaling quickly to accelerate student learning and to decrease racial and socio-economic achievement gaps (Minkos & Gelbar, 2021). Accordingly, there is a pressing need for new models of tutor training that support their professional learning and ensure that the tutoring provided to large numbers of students is of consistently high quality (Kraft & Falken, 2021, p. 10). Specific challenges faced by novice tutors include learning to foster learning communities that require active student participation, facilitate productive mathematical discussions, and engage with students equitably (Mackiewicz & Thompson, 2018; Topping, 2000). Tutors who receive on-the-job professional support have been shown to consistently produce better tutoring results (Gordon, 2009).

In much the same way research highlights the importance of dialogic learning during classroom instruction, students' active participation in discussions is also strongly encouraged by the tutoring literature (Graesser et al., 1995; Topping, 2000). For example, the literature suggests that students should be encouraged to explain rather than be provided with an explanation by a tutor (Wittwer & Renkl, 2008). More expert tutors pose questions instead of lecturing or making assertions (Dolmans et al., 2002; Lepper & Woolverton, 2002). Rosé and colleagues (2003) found that when tutors prompted students to explain their thinking, such open-ended invitations commonly provoked students to provide wrong answers, which are particularly useful from an instructional perspective to build knowledge. When working with students individually or in small groups during a tutorial session, using discursive moves aligned with accountable talk theory to construct knowledge jointly is highly effective and is generally considered "best practice" (Jitendra et al., 2013; Moschkovich, 2004).

Use of the TalkMoves Application for Coaching

Saga expressed an organizational need to enhance the on-the-job training their tutors receive with automated and personalized feedback about their discourse, similar to that provided by the TalkMoves application. In particular, Saga leadership raised the concern that novice tutors often demonstrate their math knowledge through showing and telling rather than engaging in extended conversations with students—a problem of practice (Henriksen et al., 2017) that the TalkMoves application aims to address. Independent of our work, the Saga Connect platform records tutoring sessions for coaches to review. Saga's coaches are expected to provide feedback on these videos on a regular basis as part of the coaching cycles they facilitate with each of their tutors. The discourse-focused data analytics generated by the application were identified as one way to supplement and enhance those coaching interactions. Incorporating data based on classroom observations has been demonstrated to help coaches identify areas of need, offer more targeted guidance, set measurable goals for improvement, and monitor progress over time (Glover et al., 2019).

AI-based feedback, such as that provided by the TalkMoves application, provides information about the current state of tutor–student discourse within and across tutorial sessions. The adapted TalkMoves application supports Saga's existing structure of coaching cycles, whereby coaches first review and filter the data before sharing and discussing it with tutors. As shown in Figure 20.2 (right), the intended use of the application for Saga tutors involves (1) the coach reviewing and making sense of the feedback, (2) the coach and tutor discussing the feedback and considering appropriate instructional changes, and (3) the tutor implementing the recommended instructional changes. Because tutoring sessions where Saga tutors work remotely are already routinely recorded and made available for coaches to review, much of the necessary infrastructure for the application is already in place. However, revisions to both the underlying computational models and the front end of the application were needed.

Overview of the Adapted Application

The system development process entailed integrating the Saga and TalkMoves technical architectures and data processing pipelines, updating the automated speech recognition models, fine-tuning the existing computational models based on teacher and student talk collected in mathematics classrooms to the Saga tutoring context, embedding the updated speech and language models into the analytic engine, and benchmarking the models' performance and refining as needed (see Booth et al., 2024, for details). In addition, the visualized talk moves analytics were redesigned to support the work of coaches and tutors based on a series of co-design workshops that led to new interfaces intended specifically for coaches working with tutors (Brown & Bush, 2024).

Comparing the original TalkMoves application and the adapted application, there are two major differences with respect to the NLP and automated speech recognition models (Booth et al., 2024). The adapted application uses a Roberta transformer model, similar to the TalkMoves application, but fine-tuned on data from Saga tutoring sessions. Additionally, the model in the adapted application uses an 8-past and 1-future sentence context window (in contrast to a 7-past and 7-future sentence context window) when classifying each teacher's sentence. This context window adjustment ignores irrelevant (distal future) sentences and increases the past context to provide more predictive power, especially when dealing with context-dependent discourse moves such as *keeping students together* (Suresh et al., 2022b). By employing these techniques, the updated model shows a modest performance improvement over the original. Second, the adapted application incorporates a more advanced automated speech recognition model, using OpenAI Whisper for ASR.

Pilot of the Adapted Application

An initial pilot study of the adapted TalkMoves application was conducted in Spring 2023. The study took place in a large urban public school district in the mid-western United States that primarily comprises Title 1 schools. The sample included 11 Saga coaches and the 40 tutors they supported. The coaches participated voluntarily and were asked to use the application for either seven weeks (n=6) or four weeks (n=5). An opt-out comparison group who did not have access to the application was also included in the study, consisting of 11 Saga coaches and the 28 tutors they supported. All of the tutors worked with ninth- and/or tenth-grade high school students.

Unlike the teachers who piloted the original TalkMoves application, Saga users did not have to upload their own session recordings, as this is done automatically in the Saga Connect system. To access the feedback on a given tutorial session, coaches simply click on a link within that session to view the data analytics. Nine of the 11 coaches reported viewing at least some data for the tutors they supported. The number of page views ranged from 42 to 195, with an average of 110 page views (SD=54) by each coach over the study period, or between about four and seven page views per tutor per week.

Ten coaches completed a post-study questionnaire about their interest in and perceptions of the application. Among these coaches, all responded that they were either likely (50%) or very likely (50%) to use the application on a regular basis in the future. Concerning the feedback's perceived usefulness, the average rating was 4.3 (very useful) out of a 5-point scale, and the lowest rating a coach provided was 3 (n=1). Nine of the ten coaches reported that the feedback helped them generate coaching goals for their tutors or measure progress towards previously generated coaching goals. Furthermore, all but three of the coaches indicated that they discussed the feedback with their tutors directly.

The research team was provided access to 2,350 tutoring sessions recorded during Spring 2023 (*ns*=1,307, 1,043 from the treatment and comparison groups of tutors respectively). The updated application provided data on tutor talk moves; data on student talk moves is forthcoming but was not available to users during the pilot study. Therefore, we considered changes in discourse practice only for tutor talk moves. Table 20.4 reports the standardized coefficients for change in talk move usage by tutor group over time resulting from a linear mixed effects analysis.

Results indicated that there was a measurable and statistically significant increase (p<.05) in the use of three of the talk moves for tutors whose coaches had access to the application compared to tutors whose coaches did not. The treatment tutors increased the number of times they used the talk moves *keeping students together*, *pressing for accuracy*, and *revoicing*, as well as increasing the overall number of tutor talk moves in their tutorial sessions. The change in talk move usage for comparison tutors was not significant, as expected. Interestingly, these are the same talk moves that showed a significant increase in the original TalkMoves application pilot. Moreover, this increase took place in the relatively short time span that the application was made available to their coaches, suggesting that these particular talk moves may be the easiest for tutors (and teachers) to increase with relatively minimal effort.

Table 20.4 *Standardized coefficients for change in talk move usage by tutor group over time*

Category and Tutor Talk Moves	**Treatment Tutor Group**	**Comparison Tutor Group**
All tutor talk moves	.12*	–.00
Learning Community: Keep students together	.07*	–.01
Learning Community: Get students to relate	–.03	–.02
Content Knowledge: Restating[1]	—	—
Content Knowledge: Press for accuracy	.13*	.02
Rigorous Thinking: Revoicing	.05*	–.06
Rigorous Thinking: Press for reasoning	.04	.05

Note: [1]Restating was infrequent (<1%), so it was not included in these analyses.
*p < .05.

DISCUSSION

Potential of AI-based Feedback on Discourse from Recorded Observations

A number of recent AI-based efforts have focused on identifying and classifying discourse patterns from recorded classroom observations. This is likely due to compelling evidence that attending to discursive features is highly valuable for educators' continuous improvement efforts, especially if those features align with dialogic instruction (Lefstein et al., 2020). Although at present there are only a few deployed AI tools to support teacher learning, as discussed in this chapter, research has shown that the necessary components are feasible to develop, reliable in nature, and generally well received by users (Ogan, 2019). Fully automated applications are likely to become an increasingly common approach to providing personalized, data-driven feedback from recorded classroom observations at scale (Demszky et al., 2023b; Kelly et al., 2024, Roschelle et al., 2020; Suresh et al., 2024).

The two use cases of the TalkMoves application discussed in this chapter highlight the promise and potential of AI tools in the professional learning space. In both cases, the tool served as (1) a domain expert providing automated feedback on research-based discourse practices, (2) an application of the latest NLP models applied to interpreting complex, large-scale discursive patterns, and (3) an end-to-end system designed to enhance educators' pedagogical skills to lead discourse-rich lessons. Furthermore, the use cases show how data-driven feedback presented through relatively simple visual analytics can be useful in different educational contexts, including by experienced teachers and by paraprofessional tutors working with dedicated instructional coaches. AI-based feedback from classroom recordings offers access to information about discursive interactions at a much more detailed and nuanced level than previously possible, leading to new insights into instructional practices.

The degree to which users carefully review and appropriately interpret the data generated by an AI tool is key to its ability to impact practice; however, engaging in this kind of sensemaking process is not a trivial undertaking (Campos et al., 2021; Lefstein et al., 2020). In the case of the original TalkMoves application, data analytics were provided to individual teachers who independently accessed and viewed their feedback. Analyses from the pilot study indicate that in-service teachers perceived the application to provide useful information, supporting reflective noticing (Sherin & Dyer, 2017) of patterns in their classroom discourse for targeted and self-guided improvement (Jacobs et al., 2024). By contrast, the adapted application was intended for more collaborative and facilitated interactions between coaches and tutors, in alignment with Saga's existing structure of coaching cycles guided by an articulated set of norms and expectations (Brown & Bush, 2024).

The feedback provided by the TalkMoves application is, at its core, simply a frequency count of a set of discourse events derived from accountable talk theory. By itself, the data provided to users is agnostic and non-judgmental, serving as a record of whether certain discourse events were or were not present during a recorded observation (Kelly, 2023). Moreover, the application does not offer pedagogical suggestions or set guideposts regarding which talk moves to use, how often, and in what instructional context (e.g., when during the lesson, with which students, in conjunction with certain tasks or activities). This intentional neutrality was a design decision based on teachers' expressed concerns that an automated tool should not offer judgments regarding their instruction or set unrealistic expectations without knowledge of their instructional contexts, which could feel uncomfortable and untrustworthy (Suresh et al., 2024). At the same time, a more subjective presentation of the data is certainly possible to develop and may be welcomed by some educators. Future research is needed to explore under which conditions, if any, the application should provide more subjective guidance, recommendations, and/or evaluations, and how that information might be perceived and taken up by users.

Future Directions: Knowns, Unknowns, and Challenges

The education field is currently in the early stages of developing user-facing AI-based applications that focus on classroom discourse behaviors to promote reflection and pedagogical change. However, it is becoming increasingly well understood how to develop deep learning models that can reliably detect subtle discursive acts on par with well-trained humans. Correspondingly, we can expect to see rapid growth in such models along with their applied use in an array of tools targeting student and educator learning across wide-ranging academic domains and schooling contexts (U.S. Department of Education, 2023).

Existing AI-based tools such as the TalkMoves application provide feedback on discourse moves central to dialogic classrooms using computational models that can accurately classify these dialogue acts. Future efforts will likely attend to even more complex discourse patterns, such as those that are multilingual (e.g., translanguaging) and multimodal (e.g., nonverbal communication) as more sophisticated AI methodologies are developed. In addition, applications that support dialogic teaching will need to move beyond providing feedback on discrete discourse moves and attend more holistically to the nature of classroom interactions at the intersection of discourse and socio-cultural norms, relational and affective patterns, and disciplinary content.

The TalkMoves application demonstrates the potential of modern natural language processing techniques to help improve educational discourse, but it also surfaces a number of challenges for the field. For example, many important dialogic moves occur very infrequently (Park et al., 2017), and when combined with the challenges of collecting and annotating classroom data, yield datasets with relatively few available inputs for model training (Suresh et al., 2019). Any given tool is unlikely to offer a robust, fully unbiased, "one size fits all" solution, and most will require customization for different usage scenarios to meet the needs of particular educational communities (Jensen, B. et al., 2021). Thus, we infer that an assortment of tools performing different functions for specific audiences, developed through rigorous experimentation and evaluation processes, will be necessary to make equitable and detectable progress at scale.

All of this innovation begs the question: What specifically does progress towards improved teaching and learning look like, and how can we ensure that AI-based tools do not cause unintended harm? We propose that developers and evaluators carefully consider and empirically address the following questions as they design and deploy automated tools that produce data-driven feedback derived from classroom recordings:

- What is the purpose or goal of the feedback?
- Is the feedback based on research or theory on best practices?
- Might the feedback or its presentation be construed as overly judgmental?
- What are teachers' (or other educators') perceptions of the utility of feedback?
- How should feedback be used for professional learning?
- What are the intended outcomes, for both teachers and their students?
- Does the tool help to professionalize teachers?
- What might harmful outcomes look like, and how can they be mitigated?

CONCLUSION

The Importance of Human–Technology Partnerships

Human–human relationships are central to the vast majority of educational endeavors, including those that take place in traditional classroom settings, during tutoring sessions, and in other environments that involve learning through collaboration with others. Digital tools offer the unique opportunity to participate in a human–technology partnership, which will require new parameters, norms, and mindsets for productive engagement and maximum impact (Molenaar, 2022). AI-based tools hold a great deal of promise to support teacher learning in ways that were not previously possible, with relatively low costs. Two use cases examined in this chapter involved AI-based tools to rapidly quantify large amounts of classroom data and generate discourse-related data analytics that educators found useful and that led to changes in their discussions with students.

Similar to their relationships with other humans, it is essential that people perceive their relationship with technology, and in particular with AI, as trustworthy in order for automated feedback to be accepted and willingly acted upon. AI tools should offer computational assistance that "augments and enriches, rather than replaces people's intellectual work" (Heer, 2019, p. 1844). Tools that involve recording classrooms are at particular risk of being viewed as evaluative, intrusive, judgmental, and surveillant, thereby limiting teachers' receptiveness to using data-driven feedback as a catalyst for change (Madaio et al., 2021). Engaging in purposeful and intensive co-design work can help to mitigate these risks and ensure that the resultant tools will be perceived as useful, relevant, and ethical to end users (Lin & van Brummelen, 2021; Tuomi, 2020). Numerous studies have shown that participating in co-design enables teachers to envision and bring about new forms of teaching and learning, while appropriately centering their needs and yielding innovations that are meaningful to their everyday practice (Holstein et al., 2019; Severance et al., 2016; Voogt et al., 2011). In particular, the co-design process helps potential resistances and tensions to emerge and be identified early in the development stages and then mitigated with the appropriate design choices (Leary et al., 2016).

Using AI to Support a Changing Educational Landscape

Many would argue that classroom instruction has not substantively changed over the last several decades (Gallimore & Santagata, 2006), with the implementation of rich classroom discourse practices remaining particularly elusive. At the same time, the educational landscape most certainly has experienced dramatic shifts (National Academies of Sciences, Engineering, and Medicine, 2020), including a more diverse student population, new educational standards, and, most recently, an influx of paraprofessional tutors working with students during the school day (Kraft & Falken, 2021). Automated tools used for the observation and analysis of classroom instruction can and should be positioned to productively support these changes, attending carefully to the needs of diverse students, as well as educators working under a variety of circumstances.

Tools that provide concrete feedback on discourse patterns are poised to make a profound contribution "in pursuit of the large-scale improvement of teaching" (Correnti et al., 2015, p. 303), particularly when they offer educators opportunities for noticing, reflection, and knowledge construction related to their everyday practices (Chiu et al., 2022). Major shifts in instruction—such as those necessitated by dialogic teaching that centers inclusive student participation—are likely to require ongoing critical reflection by educators about their practice (Camburn & Han, 2015; Larrivee, 2000). Automated AI-driven feedback is uniquely positioned to encourage this type of highly ambitious shift by using information gleaned from recorded instructional observations to direct attention and motivate changes in behavior.

AUTHORS' NOTE

The TalkMoves application development and research described in this chapter were funded by the National Science Foundation grants 1837986 and 1920510 and the Learning Engineering Virtual Institute (LEVI). Any opinions, findings, conclusions, or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of the National Science Foundation or the LEVI program.

NOTE

1. Processing speeds vary, but typically feedback is accessible shortly after a recording is uploaded (i.e., within 30–60 seconds).

REFERENCES

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

Applebee, A. N., Langer, J. A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*, *40*(3), 685–730.

Balyan, R., Arner, T., Taylor, K., Shin, J., Banawan, M., Leite, W. L., & McNamara, D. S. (2022). Modeling one-on-one online tutoring discourse using an accountable talk framework. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 477–483). Durham, United Kingdom. https://doi.org/10.5281/zenodo.6852936

Böheim, R., Schnitzler, K., Gröschner, A., Weil, M., Knogler, M., Schindler, A. K., … & Seidel, T. (2021). How changes in teachers' dialogic discourse practice relate to changes in students' activation, motivation and cognitive engagement. *Learning, Culture and Social Interaction*, *28*, 100450. https://doi.org/10.1016/j.lcsi.2020.100450

Bokhove, C., & Downey, C. (2018). Automated generation of "good enough" transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations*, *11*(2), 205979911879074. doi:10.1177/2059799118790743

Booth, B. M., Jacobs, J., Bush, J., Milne, B., Fischaber, T., & D'Mello, S. K. (2024). Human-tutor coaching technology (HTCT): Automated discourse analytics in a coached tutoring model. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 725–735). Kyoto, Japan. https://doi.org/10.1145/3636555.3636937

Brown, C., & Bush, J. (2024). Co-designing an AI tool to support discourse based math instruction in a high-dosage tutoring context. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning* (pp. 277–280). International Society of the Learning Sciences.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better.* Harvard Education Press.

Calcagni, E., Ahmed, F., Trigo-Clapés, A. L., Kershner, R., & Hennessy, S. (2023). Developing dialogic classroom practices through supporting professional agency: Teachers' experiences of using the T-SEDA practitioner-led inquiry approach. *Teaching and Teacher Education*, *126*, 104067. https://doi.org/10.1016/j.tate.2023.104067

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

Campos, F. C., Ahn, J., DiGiacomo, D. K., Nguyen, H., & Hays, M. (2021). Making sense of sensemaking: Understanding how K-12 teachers and coaches react to visual analytics. *Journal of Learning Analytics*, *8*(3), 60–80. https://doi.org/10.18608/jla.2021.7113

Candela, A. G., Boston, M. D., & Dixon, J. K. (2020). Discourse actions to promote student access. *Mathematics Teacher: Learning and Teaching PK-12*, *113*(4), 266–277. https://doi.org/10.5951/MTLT.2019.0009

Cazden, C. B. (2003). Classroom discourse. In C. B. Cazden & S. W. Beck (Eds.), *Handbook of discourse processes* (pp. 170–202). Routledge.

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. M. McLaren, & A. H. Alavi (Eds.), *Artificial intelligence in STEM education: The paradigmatic shifts in research, education, and technology* (pp. 163–173). CRC Press.

Correnti, R., Stein, M. K., Smith, M. S., Scherrer, J., McKeown, M., Greeno, J., & Ashley, K. (2015). Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. In L. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 303–320). American Educational Research Association.

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.

Davies, M., Kiemer, K., & Meissel, K. (2017). Quality talk and dialogic teaching: An examination of a professional development programme on secondary teachers' facilitation of student talk. *British Educational Research Journal*, *43*(5), 968–987. https://doi.org/10.1002/berj.3293

Demszky, D., & Liu, J. (2023). M-Powering teachers: Natural language processing powered feedback improves 1: 1 instruction and student outcomes. L@S '23: Proceedings of the 10th *ACM Conference on Learning @ Scale* (pp. 59–69). https://doi.org/10.1145/3573051.3593379

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023a). 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*, 01623737231169270. doi:10.3102/01623737231169270

Demszky, D., Bush, J. B., D'Mello, S. K., Jacobs, J., … & Wentworth, L. (2023b). Empowering educators via language technology. Stanford University. https://www.dorademszky.com/publications/26178-empowering-educators-via-language-technology

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186).

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

Dolmans, D., Gijselaers, W., Moust, J., Grave, W., Wolfhagen, I., & Vleuten, C. (2002). Trends in research on the tutor in problem-based learning: Conclusions and implications for educational practice and research. *Medical Teacher*, *24*(2), 173–180. https://doi.org/10.1080/01421590220125277

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017). Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 218–227). http://dx.doi.org/10.1145/3027385.3027417

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., … & D'Mello, S. K. (2016). Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (pp. 45–53). http://dx.doi.org/10.1145/2930238.2930250

Erickson, F. (2011). Uses of video in social research: A brief history. *International Journal of Social Research Methodology*, *14*(3), 179–189. https://doi.org/10.1080/13645579.2011.563615

Franklin, R. K., Mitchell, J. O., Walters, K. S., Livingston, B., Lineberger, M. B., Putman, C., Yarborough, R., & Karges-Bone, L. (2018). Using Swivl robotic technology in teacher education preparation: A pilot study. *TechTrends*, *62*(2), 184–189. doi:10.1007/s11528-017-0246-5

Gallimore, R., & Santagata, R. (2006). Researching teaching: The problem of studying a system resistant to change. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 11–28). American Psychological Association.

Glover, T. A., Reddy, L. A., Kurz, A., & Elliott, S. N. (2019). Use of an online platform to facilitate and investigate data-driven instructional coaching. *Assessment for Effective Intervention*, *44*(2), 95–103. https://doi.org/10.1177/1534508418811593

Gordon, E. E. (2009). 5 ways to improve tutoring programs. *Phi Delta Kappan*, *90*(6), 440–445.

Graesser, A., Person, N., & Magliano, J. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*(6), 495–522.

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., … & Steinberg, L. (2021). *Not too late: Improving academic outcomes among adolescents*. National Bureau of Economic Research. NBER Working Paper No. 28531.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487

Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, *116*(6), 1844–1850. https://doi.org/10.1073/pnas.1807184115

Hennessy, S., Calcagni, E., Leung, A., & Mercer, N. (2023). An analysis of the forms of teacher–student dialogue that are most productive for learning. *Language and Education*, *37*(2), 186–211.

Hennessy, S., & Davies, M. (2019). Teacher professional development to support classroom dialogue. In N. Mercer, R. Wegerif, & L. Major (Eds.), *The Routledge international handbook of research on dialogic education* (pp. 238–253). Routledge.

Henriksen, D., Richardson, C., & Mehta, R. (2017). Design thinking: A creative approach to educational problems of practice. *Thinking Skills and Creativity*, *26*, 140–153. https://doi.org/10.1016/j.tsc.2017.10.001

Herbel-Eisenmann, B. A., Steele, M. D., & Cirillo, M. (2013). (Developing) teacher discourse moves: A framework for professional development. *Mathematics Teacher Educator*, *1*(2), 181–196. https://www.jstor.org/stable/10.5951/mathteaceduc.1.2.0181

Hofmann, R., & Ruthven, K. (2018). Operational, interpersonal, discussional and ideational dimensions of classroom norms for dialogic practice in school mathematics. *British Educational Research Journal*, *44*(3), 496–514. doi:10.1002/berj.3444

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics*, *6*(2), 27–52. http://dx.doi.org/10.18608/jla.2019.62.3

Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education*, *43*(3), 325–356. https://doi.org/10.1080/0305764X.2013.786024

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Institute of Education Sciences. (2022). School pulse panel. https://ies.ed.gov/schoolsurvey/spp/

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

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *122*. https://doi.org/10.1016/j.tate.2022.103631

Jensen, B., Valdés, G., & Gallimore, R. (2021). Teachers learning to implement equitable classroom talk. *Educational Researcher*, *50*(8), 546–556. doi:10.3102/0013189X211014859

Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3313831.3376418

Jensen, E., Pugh, S. L., & D'Mello, S. K. (2021). A deep transfer learning approach to modeling teacher discourse in the classroom. In *Proceedings of the 11th International Learning Analytics and Knowledge Conference* (pp. 302–312). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448168

Jitendra, A. K., Rodriguez, M., Kanive, R., Huang, J. P., Church, C., Corroy, K. A., & Zaslofsky, A. (2013). Impact of small-group tutoring interventions on the mathematical problem solving and achievement of third-grade students with mathematics difficulties. *Learning Disability Quarterly*, *36*(1), 21–35. https://doi.org/10.1177/0731948712457561

Kelly, S. (2008). Race, social class, and student engagement in middle school English classrooms. *Social Science Research*, *37*(2), 434–448.

Kelly, S. (2023). Agnosticism in instructional observation systems. *Education Policy Analysis Archives*, *31*. 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*(5), 306–314.

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. C. (2020). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, *28*, 62. https://doi.org/10.14507/epaa.28.5012

Kelly, S., Guner, G., Hunkins, N., & D'Mello, S. K. (2024). High school English teachers reflect on their talk: A study of response to automated feedback with the Teacher Talk Tool. *International Journal of Artificial Intelligence in Education*, 1–35. https://doi.org/10.1007/s40593-024-00417-x

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*(7), 451–464. http://dx.doi.org/10.3102/0013189X18785613

Kraft, M. A., & Falken, G. T. (2021). A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*, *7(1)*, 1–21. https://doi.org/10.1177/23328584211042858

Larrivee, B. (2000). Transforming teaching practice: Becoming the critically reflective teacher. *Reflective Practice*, *1*(3), 293–307.

Leary, H., Severance, S., Penuel, W. R., Quigley, D., Sumner, T., & Devaul, H. (2016). Designing a deeply digital science curriculum: Supporting teacher learning and implementation with organizing technologies. *Journal of Science Teacher Education*, *27*, 61–77. https://doi.org/10.1007/s10972-016-9452-9

Lefstein, A., & Snell, J. (2013). *Better than best practice: Developing teaching and learning through dialogue*. Routledge.

Lefstein, A., Vedder-Weiss, D., & Segal, A. (2020). Relocating research on teacher learning: Toward pedagogically productive talk. *Educational Researcher*, *49*(5), 360–368. https://doi.org/10.3102/0013189X20922998

Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135–158). Academic Press.

Lin, P., & van Brummelen, J. (2021). Engaging teachers to co-design integrated AI curriculum for K-12 classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Yokohama, Japan. https://doi.org/10.1145/3411764.3445377

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*, *43*(4), 587–614. doi:10.3102/01623737211009267

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lyle, S. (2008). Dialogic teaching: Discussing theoretical contexts and reviewing evidence from classroom practice. *Language and Education*, *22*(3), 222–240. https://doi.org/10.1080/09500780802152499

Mackiewicz, J., & Thompson, I. K. (2018). *Talk about writing: The tutoring strategies of experienced writing center tutors*. Routledge.

Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Román, E. (2021). *Beyond fairness: Structural (in)justice lenses on AI for education*. Routledge. arXiv preprint arXiv:2105.08847.

Mercer, N., Wegerif, R., & Major, L. (Eds.). (2019). *The Routledge international handbook of research on dialogic education*. Routledge.

Michaels, S., O'Connor, M. C., Hall, M. W., & Resnick, L. B. (2010). *Accountable talk sourcebook: For classroom conversation that works*. University of Pittsburgh Institute for Learning.

Minkos, M. L., & Gelbar, N. W. (2021). Considerations for educators in supporting student learning in the midst of COVID-19. *Psychology in the Schools*, *58*(2), 416–426. https://doi.org/10.1002/pits.22454

Molenaar, I. (2022). Towards hybrid human–AI learning technologies. *European Journal of Education*, *57*(4), 632–645. doi:10.1111/ejed.12527

Moschkovich, J. (2002). A situated and sociocultural perspective on bilingual mathematics learners. *Mathematical Thinking and Learning*, *4*(2–3), 189–212. https://doi.org/10.1207/S15327833MTL04023_5

Moschkovich, J. (2004). Appropriating mathematical practices: A case study of learning to use and explore functions through interaction with a tutor. *Educational Studies in Mathematics*, *55*(1), 49–80.

National Academies of Sciences, Engineering, and Medicine. (2020). *Changing expectations for the K–12 teacher workforce: Policies, preservice education, pofessional development, and the workplace*. The National Academies Press. https://doi.org/10.17226/25603

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. NCTM.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Authors.

National Research Council. (2012). *A framework for K12 science education: Practices, crosscutting concepts and core ideas*. National Research Council.

Nickow, A., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on pre K-12 learning: A systematic review and meta-analysis of the experimental evidence*. National Bureau of Economic Research. NBER Working Paper Series 27476. http://www.nber.org/papers/w27476

O'Connor, C., & Michaels, S. (2019). Supporting teachers in taking up productive talk moves: The long road to professional learning at scale. *International Journal of Educational Research*, *97*, 166–175. https://doi.org/10.1016/j.ijer.2017.11.003

O'Connor, C., Michaels, S., & Chapin, S. H. (2015). "Scaling down" to explore the role of talk in learning: From district intervention to controlled classroom study. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through talk and dialogue* (pp. 111–126). American Educational Research Association.

O'Connor, M. C., & Michaels, S. (1996). Shifting participant frameworks: Orchestrating thinking practices in group discussion. In D. Hicks (Ed.), *Discourse, learning, and schooling* (pp. 63–103). Cambridge University Press.

Ogan, A. (2019). Reframing classroom sensing: Promise and peril. *Interactions*, *26*(6), 26–32.

Park, J., Michaels, S., Affolter, R., & O'Connor, C. (2017). Expanding the conversation: Traditions, research, and practice supporting academically productive classroom talk. In *Oxford Research Encyclopedia of Education*. Oxford University Press.

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: An analysis of the co-design process. *Research and Practice in Technology Enhanced Learning*, *2*(1), 51–74.

Ramos, J. L., Cattaneo, A. A., de Jong, F. P., & Espadeiro, R. G. (2022). Pedagogical models for the facilitation of teacher professional development via video-supported collaborative learning. A review of the state of the art. *Journal of Research on Technology in Education*, *54*(5), 695–718. https://doi.org/10.1080/15391523.2021.1911720

Reinholz, D. L., & Shah, N. (2018). Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. *Journal for Research in Mathematics Education*, *49*(2), 140–177. http://www.jstor.com/stable/10.5951/jresematheduc.49.2.0140

Resnick, L., Asterhan, C., & Clarke, S. (2015). *Socializing intelligence through academic talk and dialogue*. American Educational Research Association.

Resnick, L. B., Asterhan, C. S., 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.

Roschelle, J., Lester, J., & Fusco, J. (2020). *AI and the future of learning: Expert panel report*. Digital Promise. https://files.eric.ed.gov/fulltext/ED614308.pdf

Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003). The role of why questions in effective human tutoring. In *Proceedings of Artificial Intelligence in Education*, *13*, 55–62.

Severance, S., Penuel, W. R., Sumner, T., & Leary, H. (2016). Organizing for teacher agency in curricular co-design. *Journal of the Learning Sciences*, *25*(4), 531–564. https://doi.org/10.1080/10508406.2016.1207541

Sherin, M. G., & Dyer, E. B. (2017). Mathematics teachers' self-captured video and opportunities for learning. *Journal of Mathematics Teacher Education*, *20*(5), 477–495.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

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*(3), 496–521. http://dx.doi.org/10.1177/0735633120968554

Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J., & Sumner, T. (2022a). The TalkMoves Dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. Presentation at the *13th International Conference Language Resources and Evaluation Conference*. https://arxiv.org/abs/2204.09652

Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J., & Sumner, T. (2021). Using transformers to provide teachers with personalized feedback on their classroom discourse: The TalkMoves Application. Paper presented at the *Association for Advancement of Artificial Intelligence Symposium on Artificial Intelligence for K-12 Education*. https://arxiv.org/pdf/2105.07949.pdf

Suresh, A., Jacobs, J., Perkoff, M., Martin, J., & Sumner, T. (2022b). Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 71–81). Association for Computational Linguistics. 10.18653/v1/2022.bea-1.11

Suresh, A., Penuel, W., Jacobs, J., Raza, A., Martin, M., & Sumner, T. (2024). Using AI tools to provide teachers with fully automated, personalized feedback on their classroom discourse patterns. In X. Zhai & J. Krajcik (Eds.), *Uses of artificial intelligence in STEM education* (pp. 371–398). Oxford University Press. 10.1093/oso/9780198882077.003.0017

Suresh, A., Sumner, T., Jacobs, J., Foland, B., & Ward, W. (2019). Automating analysis and feedback to improve mathematics teachers' classroom discourse. In *Proceedings of the Association for Advancement of Artificial Intelligence Conference on Artificial Intelligence*, *33*(1), 9721–9728. https://doi.org/10.1609/aaai.v33i01.33019721

Tao, Y., & Chen, G. (2023). Coding schemes and analytic indicators for dialogic teaching: A systematic review of the literature. *Learning, Culture and Social Interaction*, *39*, 100702. https://doi.org/10.1016/j.lcsi.2023.100702

Teo, P. (2019). Teaching for the 21st century: A case for dialogic pedagogy. *Learning, Culture and Social Interaction*, *21*, 170–178. https://doi.org/10.1016/j.lcsi.2019.03.009

Topping, K. (2000). *Tutoring.* International Academy of Education. Educational Practices Series 5. Switzerland.

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.

Tuomi, I. (2020). *Research for CULT Committee: The use of Artificial Intelligence (AI) in education.* European Parliament, Directorate-General for Internal Policies, 2–6.

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and future of teaching and learning: Insights and recommendations.* Washington, DC. https://tech.ed.gov/files/2023/05/ai-future-of-teaching-and-learning-report.pdf

van der Lans, R. M., van de Grift, W. J., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, *50*, 88–95. https://doi.org/10.1016/j.stueduc.2016.08.001

Voogt, J., Westbroek, H., Handelzalts, A., Walraven, A., McKenney, S., Pieters, J., & De Vries, B. (2011). Teacher learning in collaborative curriculum design. *Teaching and Teacher Education*, *27*(8), 1235–1244. https://doi.org/10.1016/j.tate.2011.07.003

Wagganer, E. L. (2015). Creating math talk communities. *Teaching Children Mathematics*, *22*(4), 248–254. https://www.jstor.org/stable/10.5951/teacchilmath.22.4.0248

Walshaw, M., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research*, *78*(3), 516–551. https://doi.org/10.3102/0034654308320292

Walter, H. A. (2018). Beyond turn and talk: Creating discourse. *Teaching Children Mathematics*, *25*(3), 180–185. https://doi.org/10.5951/teacchilmath.25.3.0180

Wang, D., Shan, D., Zheng, Y., Guo, K., Chen, G., & Lu, Y. (2023). Can ChatGPT detect student talk moves in classroom discourse? A preliminary comparison with Bert. In M. Feng, T. Kaser, and P. Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 515–519). Bengaluru, India. International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115772

Webb, N. M., Franke, M. L., Ing, M., Johnson, N. C., & Zimmerman, J. (2019). The details matter in mathematics classroom dialogue. In *The Routledge international handbook of research on dialogic education* (pp. 530–546). Routledge.

Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., & Turrou, A. C. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research*, *63*, 79–93. https://doi.org/10.1016/j.ijer.2013.02.001

Wegerif, R. (2013). *Dialogic: Education for the internet age*. Routledge.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist*, *43*(1), 49–64. https://doi.org/10.1080/00461520701756420

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., … & Ahmed, A. (2020). Big bird: Transformers for longer sequences. *34th Conference on Neural Information Processing Systems*, *33*, 17283–17297. Vancouver, Canada. https://doi.org/10.48550/arXiv.2007.14062