# A Framework for Considering Exploration, Interpretation, and Confirmation During Data Analysis: Computationally Assisted Analysis of Teacher–Group Interactions

Paul Hur
Freie Universität Berlin
Berlin, Germany
k.hur@fu-berlin.de

Chris Palaguachi
University of Illinois Urbana–Champaign
Urbana, USA
cwp5@illinois.edu

Nessrine Machaka
University of Illinois Urbana–Champaign
Urbana, USA
machaka2@illinois.edu

Christina Krist
Stanford University
Stanford, USA
stinakri@stanford.edu

Elizabeth B. Dyer
University of Illinois Urbana–Champaign
Urbana, USA
edyer@illinois.edu

Cynthia D'Angelo
University of Illinois Urbana–Champaign
Urbana, USA
cdangelo@illinois.edu

Nigel Bosch
University of Illinois Urbana–Champaign
Urbana, USA
pnb@illinois.edu

Education researchers increasingly analyze heterogeneous, multimodal data with computational tools. Yet, reporting rarely makes explicit who (human or computer) leads meaning-making at different points in the analysis. We introduce a framework for analytic agency that distinguishes three stages, exploration, interpretation, and confirmation, and classifies each as primarily human- or computer-led, as considering stage-level leadership can clarify assumptions in analysis. We demonstrate the framework in a multimodal case study of teacher–student group interactions in high school mathematics classrooms. Using 15 classroom videos from three teachers, we selected 21 student groups and developed a pose-based detector that flags interactions. The pipeline aligned group-level audio and word-level transcripts to each detected window and computed acoustic/prosodic features and large-language-model indicators for question-asking, confusion, help-seeking, and math talk. Across the corpus, the detector surfaced 317 interaction events (*M* = 15.10 per group, *SD* = 12.42;

mean duration = 32.73s). We compared before, during, and after segments using paired tests and mixed-effects models. Naturally, results for mixed-effects models showed significant shifts in keypoints before-to-during and before-to-after for those emphasized in the detection approach, while audio features showed no significant changes. One transcript indicator, confusion, decreased after interactions ($\beta = -0.061$, $p = .049$). The pipeline showed preferences for spatial co-presence rather than interaction discourse change, which illustrates how leadership in exploration shaped what became detectable and, consequently, how interpretation proceeded. In the paper's conclusion, we outline hybrid, iterative variants and discuss limitations. Making stage-level agency explicit can help researchers align methodological choices with theoretical aims and produce more transparent, auditable analyses of complex classroom data.

**Keywords:** classroom video data, student group interactions, hybrid analysis, computational grounded theory

## 1. INTRODUCTION

Increasingly, education researchers are creating, mixing, and employing research methods in ways as diverse as the data they are analyzing (Romero & Ventura, 2020; Yoon & Hmelo-Silver, 2017), and this methodological diversity requires a greater emphasis on making analytical choices transparent and defensible (Bergner et al., 2018). The challenge of analytic agency (i.e., the primary source of inquiry or decision-making) is rooted in the diversity of the student data, as researchers regularly contend with different modalities such as written classroom artifacts, video, audio, or action logs (e.g., Cohn et al., 2024; Gabbay & Cohen, 2022) from a variety of contexts, including in-person classrooms, online, or immersive learning environments (e.g., virtual reality, educational games; Mejia-Domenzain et al., 2024; Shute, 2011). Further increasing the complexity are the granularity and scale of these data, which can range from fine-grained physiological sensor readings and computer logs to massive datasets generated by thousands of students (e.g., Mills et al., 2021; Choi et al., 2020).

Online learning and other computer-based learning environments can collect student action log data, or abstracted, system-level traces (e.g., clicks, submissions), which represent student interactions with the system. These log data sources are common in educational data mining and related fields, as their quantitative nature is often a natural fit for computational methods and machine learning models to predict and understand student outcomes and behaviors. Insights derived from these analyses are often the basis of the underlying mechanisms of intelligent tutoring systems, which can automatically personalize and adapt individual learning experiences, and are perhaps the most well-known and strongly validated form of AI in education (D'Mello & Graesser, 2023). Moreover, a survey of methods in the field has found that most traditional data mining techniques, such as classification, clustering, and prediction, have been maturely and successfully applied in education (Romero & Ventura, 2020). Yet, the perceived objectivity and the reliance on these computational pipelines make the call for transparent and defensible methods more crucial to discuss (e.g., Baker et al., 2024; Bergner et al., 2018), as the researcher's own analytical choices and assumptions become simultaneously more influential and less immediately apparent. While computational methods can identify patterns at scale, there is value in due diligence from the researcher to be aware of the implications of the methodological choices carried out by non-human-driven decisions, as not doing so can move the researcher further away from the benefits of human-guided processes of inquiry, such as grounded theory, which have long been regarded as a way to add human context to the meaning of findings (Charmaz, 2014).

Computational research pipelines can create a second, unexamined “black box”, which is not just around the AI models and computational tools, but around the researcher’s own interpretive inquiry process (Wenskovitch & North, 2020). This issue is beyond a lack of consistent reporting practices; rather, it is a communication challenge between two distinct learning entities. While a researcher operates with high-level theoretical constructs, evolving goals, and rich domain and contextual knowledge, computational tools operate on low-level statistical features and mathematical formulas. This creates a juxtaposition where a researcher may decide on purely computational or purely qualitative approaches, or, in cases of mixed methods, choose to combine approaches in the absence of many established perspectives. As a result, the researcher’s judgments, such as which features to engineer from theory, how to label an ambiguous data point, or when to discard an overfitted model that is accurate but theoretically irrelevant, are treated as static configurations or external commands, rather than as dynamic, meaningful signals the system could learn from. There is thus a need for discussions about human-led or computer-led decision-making in various analytical processes in research.

The lack of a common language to describe the researcher’s role can lead to what seems to be a false dichotomy between what are perceived as human-driven qualitative methods and automated computational ones. In practice, this distinction is rarely straightforward. Machine learning pipelines are guided by human theoretical choices, while qualitative analyses are often facilitated, at least in part, by computational tools. The issue, therefore, is not a failure to choose the right paradigm, but the absence of a shared vocabulary to describe the internal sequence of human-led and computer-led steps within a given method. Without this vocabulary, researchers cannot easily compare the point of analytical agency across different approaches, nor can they be transparent about where their own judgments shaped the analysis versus where computational tools took the lead.

In this paper, we introduce a framework that provides this necessary vocabulary, which we propose could lead to a more precise and transparent discussion of methodological design. We distinguish the data analysis process into three stages: (1) exploration, (2) interpretation, and (3) confirmation, and propose classifying methods based on whether each stage is primarily human-led or computer-led. This simple typology allows us to consider the entity that more heavily influences the goals in the steps across various methods within a common language. Our framework clarifies how hybrid methods like computational grounded theory (Nelson, 2020) create human–computer partnerships of inquiry that effectively delegate primary decision-making influence on the human researcher or automated computational tools based on the stage of data analysis. We demonstrate this framework through a case study using audio and video data of high school mathematics classrooms to identify and explore teacher–student group interactions. By way of describing and discussing the case study and its results, we convey the importance of being conscious of the analytical leadership of humans and computational tools (including AI) for data exploration and interpretation. Although human judgment occurred, we show that these stages in the case study were in practice, computer-led, as the computational outputs and tool defaults established the effective decision boundaries. We contribute to educational data mining by proposing a simple framework for making analytic agency explicit across common data analysis stages, and by demonstrating its use through a multimodal pipeline for detecting and analyzing teacher–student group interactions in classroom video data. This led to the central framing question of our study:

> *How can a framework that distinguishes between human-led and computer-led stages of analysis help researchers more deliberately and transparently navigate the design of methodological approaches?*

The rest of the paper is organized as follows. The next section introduces our simple analytics agency framework classified across exploration, interpretation, and confirmation, then proceeds to background sections which provide context to the discussions in the related frameworks around analytics workflows and agency, including human-in-the-loop (HITL) and computational grounded theory, and covers the multimodal classroom data research and tools about the computational research landscape for our case study. This will be followed by the methods section to describe our data, feature extraction, and data processing, as well as our approach to teacher–student group detection. In the results section, we describe the findings of our computer-led data analysis. Within the discussion, we revisit the simple typology as a table with common research methods, and we reflect on the case study within the framework, as well as a thought experiment about various counterfactual analytical leadership scenarios (e.g., human-led exploration, computer-led interpretation) to understand how it can bring about hidden assumptions about analytics agency. Lastly, we conclude with a summary of our limitations and a discussion of future directions. In sum, we intend our case study to be one example within the framework and not necessarily the "correct" way or a method recommendation for sensemaking or inquiry regarding classroom video analysis, but to be viewed as an example for researchers to be more deliberate about and aware of where analytical agency lies when working with AI and other computational tools.

## 1.1. Overview of The Analytics Agency Framework

We propose a simple framework for analytic agency that treats data analysis as proceeding through three stages: exploration (how data exist, and are transformed or selected), interpretation (how patterns are made sense of, labelled, or theorized), and confirmation (how findings are tested, validated, or generalized). For each stage, we focus on stage leadership, whether the consequential decision-making lies with the human researcher or with computational tools. In other words, the framework asks, at each stage, who or what most strongly shapes which questions are asked, which patterns are noticed, and which results are treated as meaningful.

To make these lenses concrete, we can classify analytic workflows by who (human or computer) leads each of the three stages of data analysis: exploration, interpretation, and confirmation. We have situated common methods and our case study (in the last row) within that space in Table 1. The entries are not prescriptive; we have placed them based on the default locus of decision-making most practitioners would recognize. We use "N/A" where a stage is typically not formalized in the method, and "None" where interpretation is effectively outsourced (e.g., to pre-existing labels) rather than performed within the pipeline. Reading methods this way turns positionality into specific, stage leadership choices that can be named and compared.

Table 1: Classifying common methods via human-led and computer-led stages of analysis. Note that Human and Computer labels mean primarily led by that entity, rather than necessarily without any contribution from the other.

| Exploration | Interpretation | Confirmation | Method |
|---|---|---|---|
| Human | N/A | Computer | Statistical hypothesis testing |
| Human | Human | N/A | Qualitative analysis |
| Computer | None | Computer | Supervised machine learning |
| Computer | Human | N/A | Unsupervised machine learning |
| Computer | Human | Computer | Computational grounded theory |
| Human | Computer | N/A | Manual coding with computer-assisted descriptive summaries |
| Computer | Computer | N/A | Our presented case study |

We return to this table in our discussion, when interpreting our case study results and considering counterfactual case study designs (e.g., human-led exploration with computer-led interpretation) to illustrate how different leadership choices could produce different forms of evidence and insight from the same data.

### 1.2. Analytic Workflows and Agency: HITL/Interactive ML to Computational Grounded Theory

Work on human–AI collaboration in education has largely shown how people can be inserted into computational pipelines, through labeling, feature construction, or post-hoc interpretation, without explicitly naming who leads meaning-making at different points in the analysis (Lin & Koedinger, 2017; González-Brenes & Mostow, 2012). In our work, we use analytic agency to refer to the entity (human or computer) that primarily sets goals and makes consequential choices within a stage of analysis. Emphasizing agency aligns with calls in computational research for transparency about analytic choices and their assumptions (Ouhaichi et al., 2024; Bergner et al., 2018; Baker et al., 2024) and with concerns that mixed-modal, large-scale data can obscure where interpretation occurs (Ochoa & Worsley, 2016).

Human-in-the-loop (HITL) and interactive machine learning place opportunities for human involvement inside computer-heavy workflows (e.g., Lu et al., 2024, Dragut et al., 2021). Often, points for human involvement include curating and labeling data, engineering features from theory, steering models (e.g., constraints, hyperparameters), and using explainable-AI tools to interrogate model behavior. These approaches improve practicality and reliability in learning analytics, where domain knowledge matters (e.g., Venkatesha et al., 2025; Rajarathinam et al., 2025). However, the emphasis is on interaction opportunities rather than stage leadership. We refer to stage leadership when we talk about which entity effectively leads a specific stage (exploration, interpretation, or confirmation) by shaping what questions are asked, what patterns are observed, and which results are treated as consequential. If algorithmic defaults determine what is surfaced during exploration (e.g., which segments, clusters, or features are even considered), later human labeling or explanation occurs within a machine-defined space (Sandvig, 2014).

Computational grounded theory (CGT) makes the division of labor more explicit by assigning large-scale pattern finding to computation and theory building/refinement to researchers, typically via an iterative cycle of *pattern finding → pattern refinement → validation* (Nelson, 2020). CGT has enabled qualitative inquiry at scale (especially for text) while keeping human interpretation central (Salvi & Bosch, 2025). Yet, leadership is often implied. Preprocessing and clustering criteria underlying computational processing constrain what becomes available for human theorizing. In our paper, we adopt three stages common to most empirical studies –

exploration, interpretation, and confirmation – and treat each stage as primarily human- or computer-led, to make leadership comparable across methods.

### 1.3. Multimodal Learning Analytics for Classroom Event Detection

Recent advances in computer vision and speech processing have made it possible to determine classroom behaviors and events at scale using video and audio data. Off-the-shelf pose estimation can extract these characteristics from video, such as OpenPose or VIBE (Cao et al., 2021; Kocabas et al., 2020), and modern automatic speech recognition can produce clean transcripts from audio recordings (e.g., Radford et al., 2023; Bredin et al., 2020). Additionally, audio processing tools can analyze prosody and voice quality over time (Boersma & Van Heuven, 2001; Eyben et al., 2010). These tools provide low-level representations that are well-suited for exploration of their original data. Our goal in the present work is to leverage this maturity for event detection and to keep the pipeline methodologically meaningful.

Movement signals are highly informative for learning research. Pose information, for instance, can be used to infer student engagement levels by analyzing body posture and orientation (Zhao et al., 2021; Vieira et al., 2021). Furthermore, finer-grained analyses of gestures and shifts in posture can signal key pedagogical moments, such as when a student or teacher is explaining their reasoning about mathematics (Alibali & Nathan, 2012). While real-world classroom data presents challenges, such as students obscuring one another, tracking key body points within defined group areas can provide stable data for analysis. This has allowed for the development of methods to flag moments of potential interest for researchers, such as with EduSense (Ahuja et al., 2019).

Speech can provide another layer of data to classroom learning contexts by capturing the content and delivery of classroom talk. For example, automated transcripts have been used in large-scale analysis of texts to understand and be used for creating better instructional support (Whitehill & LoCasale-Crouch, 2023). Beyond the words themselves, non-semantic acoustic features like pitch, volume, and the rhythm of speech have been used to determine student engagement, confidence, or uncertainty (Singh et al., 2025; Kumar et al., 2024; Frankel & Brownstein, 2016). In our paper, we combine these three modalities to develop a computational pipeline to automatically detect and analyze teacher-student group interactions in real-world classroom settings.

## 2. Method

Our case study focused on essential moments for understanding responsive pedagogy: the interactions that occur when a teacher moves between small student groups. These improvised, dynamic interactions are a staple in classroom learning, and teachers often employ proactive classroom management strategies to most effectively determine how and which student groups to support (Kaendler et al., 2015). We viewed them as an ideal site for methodological inquiry as they are rich with verbal and non-verbal cues such as gestures and useful for understanding learning, but are challenging to study systematically at scale (Gonzales et al., 2019). In this section, we describe the methods used in our case study to identify and analyze teacher–student group interactions in classroom video data. We describe our data, the pose and audio feature extraction, our teacher–group interaction detection method, statistical analysis of the computationally determined clips, and a qualitative assessment of the clips based on understanding student group gaze.

## 2.1. Classroom Video Data and Group-Level Audio Data

The data were collected for a research project that took place in 2013 and 2014 in high school mathematics classrooms in the United States. The goal of the research project was to understand high school mathematics teachers' pedagogical practices and to consider analyzing footage captured via point-of-view camera angles. Additionally, fixed-angle cameras were placed at various corners, typically at the front of the room, to capture the full classroom, including student interactions with the teacher as well as with each other. These fixed-angle videos were recorded in 1080p (1920×1080 pixels) resolution at 30 frames per second, and depending on the camera used for the capture, the videos had 120-degree and 130-degree fields of view. A total of 15 of these roughly 90-minute classroom videos (examples in Figure 1) were included for our analysis across three different teachers: six for Teacher 1, five for Teacher 2, and four for Teacher 3.

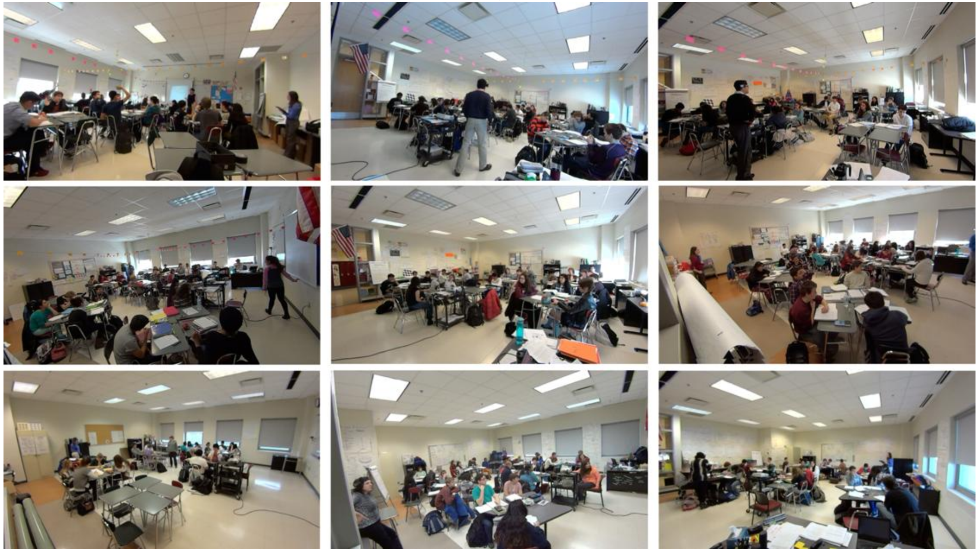


Figure 1: Example classrooms from the 15 mathematics videos analyzed.

All classrooms in the videos had students seated in small groups of mostly three or four students, though some attendance variations led to some groups having only two students. We selected 21 student groups from the 15 classroom videos based on preliminary analysis of watching clips throughout points within the class period. Specifically, we selected student groups that were mostly unobscured by other non-group individuals or classroom furniture and discrete (i.e., having relative separation from other groups). All student groups selected for our analysis had a minimum of three students.

While audio data were captured from the cameras' microphones, more precise group-level audio were also captured by a dedicated audio recorder placed at the center of each of the small group tables. As these recorders were placed closer in proximity to the student group individuals, the resulting audio were able to capture clearer voices and conversations of the group during the class period. Ambient and nearby noises (e.g., non-target group conversations) were inevitably picked up by these microphones but were minimized compared to the cameras' microphones.

## 2.2. Pose and Audio Feature Extraction

We carried out feature extraction processes to obtain three low-level characteristics from the data: pose information, acoustic (i.e., physical properties such as energy or pitch), and prosodic (i.e., rhythmic properties such as stress and contour) audio feature representations, and audio conversation transcripts. To obtain pose information, we first processed all videos using OpenPose (Cao et al., 2021), an open-source computer vision tool that extracts detected individuals' skeletal low-level pose information in the form of ordered keypoints at the videos' frame-level. In other words, each frame in the video is independently analyzed by OpenPose to detect and estimate the individuals' pose orientations, which OpenPose can output as a JSON file with frame numbers and those detected individuals' keypoints as X and Y coordinate values on the video's resolution. While OpenPose offers a few configuration options to change the keypoint granularity (e.g., 15, 18, 25 body keypoints, 70 face keypoints), we chose the 25-keypoint body configuration to maintain a balance between meaningful granularity of body poses and excluding extraneous keypoints, such as detailed face keypoints, which were infeasible due to the individuals' distance from the camera.

OpenPose does not have functionality to track the detected individuals between the frames, i.e., to determine whether a person detected in a particular position is the same person detected in a similar position in the next frame. Thus, we applied a previously developed open-source postprocessing method based on calculating Euclidean distance of the keypoint values to determine closest matches (Hur & Bosch, 2022). This process results in assigning IDs to the raw keypoints, allowing us to track the individuals over time. A core feature of this postprocessing process allows for designating the boundary box region of the group of interest via four values: $X_{left}$, $Y_{top}$, $X_{right}$, and $Y_{bottom}$. We used coordinate pixel values within the video resolution (1920×1080) to determine appropriate boundaries around the target group, while attempting to minimize other individuals within it to reduce noise in the data. By outlining the target group through this process (e.g., Figure 2), we were able to restrict and consider the data of each target group. Nine student groups were selected from Teacher 1's videos, seven from Teacher 2, and five from Teacher 3.

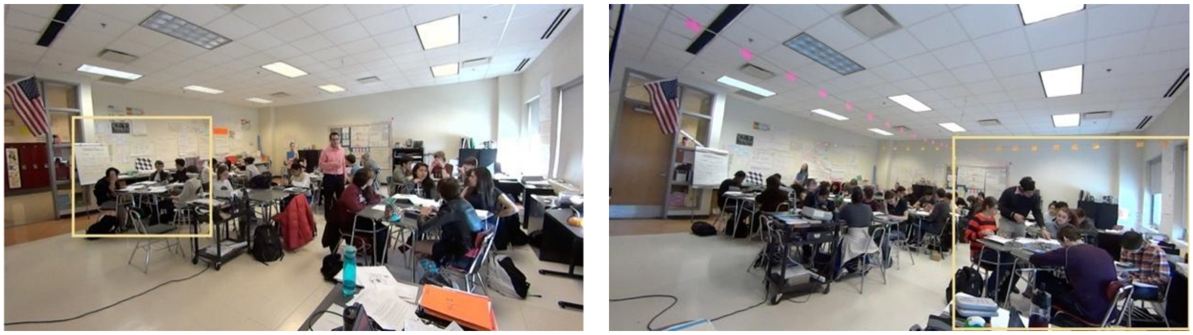


Figure 2: Example visualizations of the boundary box regions in yellow to track and restrict the pose data to that of the target student group.

To process audio data, we used openSMILE (Eyben et al., 2010), which is an open-source project that extracts a wide range of features for audio signals. For our analysis, we used the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) configuration (Eyben et al., 2016). This feature set was selected as it provided a standardized and compact summary of acoustic (e.g., loudness, shimmer) and prosodic (e.g., contour, speech rate) information relevant to social signal processing and affective computing (e.g., Xue et al., 2019, Haider et al., 2021). We first synchronized the group-level audio files to their respective videos (i.e., resolving the offsets resulting from original data collection comprising video cameras and table-placed microphones) to match the student group audio to the pose data, and applied this process to all group-level audio recordings. The 88 features of the eGeMAPS set are generated by first calculating a smaller set of low-level descriptors, such as pitch, loudness, and spectral characteristics, over short audio frames. We used the default parameters, which analyze a frame length of 25ms extracted every 10ms. Subsequently, openSMILE also applies various statistical functionals (e.g., mean, standard deviation, percentiles) to these low-level descriptors' contours across the entire audio file.

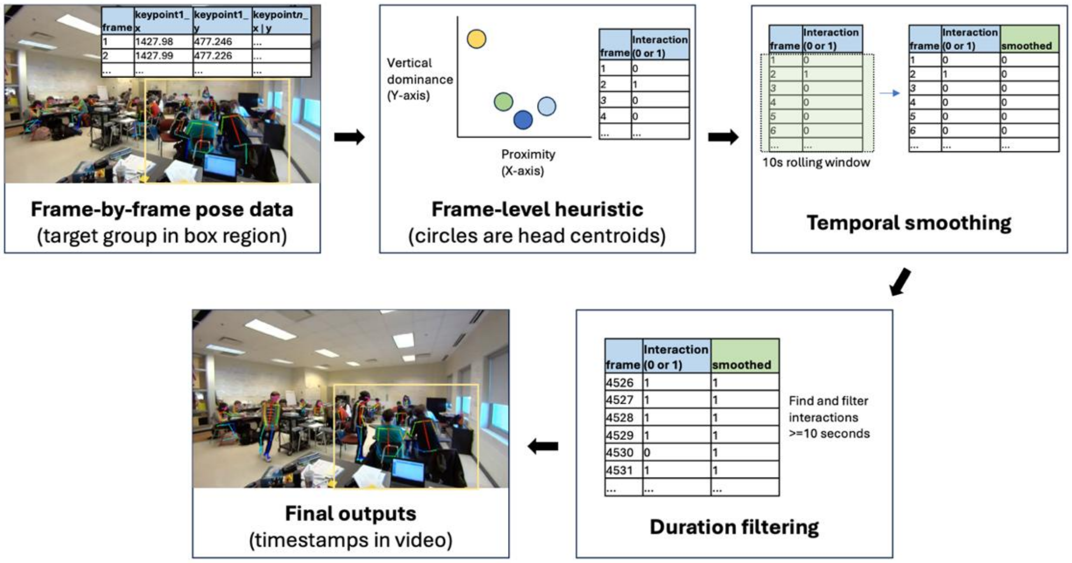
The group-level audio files were also used to extract text transcripts. We used WhisperX (Bain et al., 2023), a system that enhances the transcription capabilities of OpenAI's open-source Whisper automated transcription model (Radford et al., 2023). While Whisper provides highly accurate transcription, its default timestamps are for entire utterances. Our analysis required finer-grained word-level timestamps to more accurately align verbal events with the non-verbal pose data. Furthermore, this system uses the pyannote.audio voice activity detection model (Bredin et al., 2020) to preprocess the long classroom audio files for human speech, which optimizes the input and reduces the repetition errors that can occur when transcribing long classroom audio files. Our choice with WhisperX is further supported by research, which has been rigorously evaluated as having reasonable performance in correctly transcribing error-included spontaneous speech instances (74% correct for word error scenarios and 83% correct for sound error scenarios) (Alderete et al., 2025). Moreover, researchers analyzing classroom dialogue in naturalistic small-group settings have been using modified versions of Whisper, such as WhisperX for extracting speech-related features (Yin et al., 2025; Sivakumaran et al., 2025; French et al., 2025). While we did not carry out manual cleaning of the resulting files, we compared several Whisper and WhisperX outputs and found that WhisperX's versions consistently output more readable transcriptions with fewer noticeable errors (e.g., some Whisper transcriptions showed "Yeah." hundreds of times in a row).

## 2.3. Teacher–Group Interaction Detection Method

To systematically identify moments of teacher–student group interaction within the ~90-minute classroom videos, we developed and applied a computational method to analyze the tracked pose data. Our goal was to generate a set of timestamped video clips for each of the 21 student groups where a teacher was likely engaging with the group. Our approach first identified potential interaction frames based on heuristics and then applied temporal smoothing to ensure the resulting clips are stable for analysis. This process transformed the frame-by-frame pose data into discrete interactions.

The core of our method was a heuristic that defined a teacher's interaction based on their physical orientation relative to the seated students in the group. For each frame, we first calculated a head position for every detected person using a weighted centroid of keypoints around the head and neck (specifically, nose, neck, eyes, and ears). This weighted approach was done to minimize jitter from minor head movements. This heuristic was based on a simple observation from the videos that the teacher was much more likely to be standing, especially around the students, than sitting. Thus, interaction was hypothesized to occur when one individual (i.e., the presumed teacher) was both vertically dominant – as determined by the Y-axis values – and in proximity, via the X-axis values, to the other individuals (i.e., the students). A frame was flagged as a potential interaction only if a single person's head centroid (the presumed teacher's) was significantly higher than all others and was within proximity to the student group. The interaction detection heuristic was intentionally kept simple and interpretable. We operationalized classroom patterns of teachers typically standing while students remain seated and used pose height and spatial distance as clues to likely teacher–group interactions. Although this may be a simple heuristic for teacher–group interactions, a similar method of detecting the teacher at the group was previously used on this dataset for qualitative research regarding teacher–group interactions (Parr, 2021). The choice aligns with our broader aim of illustrating how even relatively simple, computer-led exploration decisions can shape what becomes detectable and thus what could be interpreted as meaningful.

This classification is inherently noisy, so a temporal filtering process was necessary to isolate sustained interactions from momentary detections. To achieve this, we smoothed the binary sequence of potential interaction frames using a 10-second rolling window. This step effectively determined if the interaction is consistent over a short period. For any given moment to be considered part of a probable interaction, a high percentage of the frames within its 10-second window had to meet our heuristic criteria. Finally, in order to filter out very short interactions (e.g., someone walking by, brief fly-by encounters, detection errors), we identified continuous streaks of these confident interaction periods and retained only those that lasted for a minimum of 10 seconds, or approximately 300 frames. This resulted in a set of teacher–student group interactions video clips per group for subsequent analysis. The full process is depicted in Figure 3.



| frame | keypoint1_x | keypoint1_y | keypoint*n*_x \| y |
|---|---|---|---|
| 1 | 1427.98 | 477.246 | ... |
| 2 | 1427.99 | 477.226 | ... |
| ... | ... | ... | ... |

| frame | Interaction (0 or 1) |
|---|---|
| 1 | 0 |
| 2 | 1 |
| 3 | 0 |
| 4 | 0 |
| ... | ... |

| frame | Interaction (0 or 1) |
|---|---|
| 1 | 0 |
| 2 | 1 |
| 3 | 0 |
| 4 | 0 |
| 5 | 0 |
| 6 | 0 |
| ... | ... |

| frame | Interaction (0 or 1) | smoothed |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 1 | 0 |
| 3 | 0 | 0 |
| 4 | 0 | 0 |
| 5 | 0 | 0 |
| 6 | 0 | 0 |
| ... | ... | ... |

| frame | Interaction (0 or 1) | smoothed |
|---|---|---|
| 4526 | 1 | 1 |
| 4527 | 1 | 1 |
| 4528 | 1 | 1 |
| 4529 | 1 | 1 |
| 4530 | 0 | 1 |
| 4531 | 1 | 1 |
| ... | ... | ... |

Figure 3: A visual representation of our method for determining teacher–student group interactions, starting with the pose data and ending with timestamps of video (via frame numbers).

The detection method used the video frame number as the temporal unit, so these were converted into timestamps by calculating the analogous timestamps based on the video frame rate. With the start and end times for each teacher–student interaction clip identified, we then aligned the corresponding audio data streams. For the acoustic and prosodic features generated by openSMILE, we filtered the time-series data to include only the features within the timestamp of each event. Similarly, using the word-level timestamps from WhisperX, we collected all transcribed words that were spoken during these periods. This step ensured that for every interaction identified through pose analysis, we had a synchronized set of audio features and transcript data.

To assess the reliability and construct validity of the detection method, we conducted a ground-truth validation using stratified random sampling. We selected one full ~90-minute video from each of the three teachers, and within each video, randomly selected one student group and manually coded all "true" teacher–student group pedagogical interactions, which we defined as sustained periods (10 seconds or greater) where the teacher was actively engaging with the group. We then compared these with the human-coded timestamps against the detector's output to calculate a recall metric and analyze the causes of the undetected events.

## 2.4. Statistical Analysis of Computationally Determined Clips

We statistically analyzed each of the computationally determined clips in terms of the audio, video, and transcribed speech modalities to determine whether there were notable differences before, during, and after teacher–group interactions. Specifically, we analyzed *before vs. during* and *before vs. after* to understand whether a teacher's intervention was apparent either during the interaction itself or perhaps that the teacher's influence changed the nature of student group interactions even after the teacher's interaction was over.

### 2.4.1. Segmentation

We segmented the data into before, during, and after clips based on the duration of the teacher–group interaction event, i.e., the duration of the "during" clip, so that for each event there would be three comparable clips. Figure 4 shows how clip durations were selected so that the before/during and before/after clips would be comparable within-event (i.e., comparing the same amount of time before and during a teacher–group interaction), but clip durations were still allowed to vary across interaction events, given that interactions were of different lengths.

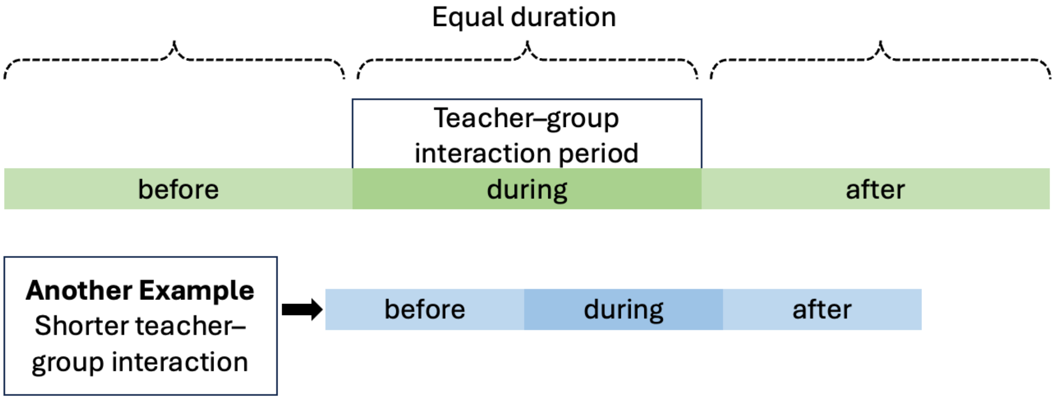

Equal duration
Teacher–group interaction period
before
during
after
Another Example
Shorter teacher–group interaction
before
during
after

Figure 4: Illustration of time segments compared for statistical analysis, where the duration was matched within each interaction event but allowed to vary across events.

### 2.4.2. Processing, aggregation, and validation

We aggregated the data within each segment to characterize the segments numerically for statistical comparison across time periods (where time period refers to segments before, during, or after a teacher–group interaction event). For video and audio modalities, we simply calculated the mean of each column within each segment. Transcripts, on the other hand, were more complicated because they consisted of words in dialogue turns, rather than easily aggregated numbers. We applied a large language model (specifically, Mistral-Small-3.2-24B-Instruct-2506, Mistral AI, 2025) that we ran on-premises (i.e., locally) to avoid data privacy issues related to sharing potentially identifying metadata and information in transcripts with cloud-based language model providers.

We applied a few basic language model prompts to measure the presence or absence of potentially classroom-relevant constructs, which were text features around question-asking, confusion, need for help, and math-relevant discourse. For example, the question-asking prompt was formatted as:

> *Do any of the snippets of dialogue below contain an explicit question?*
> *Respond with only Yes or No.*
>
> *[student 1 conversational turn 1]*
> *[student 2 conversational turn 1]*
> *[…additional conversational turns (range: 1-65)…]*

The other prompts replaced "contain an explicit question" with "indicate that students [placeholder]" where the placeholder was "might be confused", "need help from a teacher", or "are discussing math". Language model sampling parameters – i.e., how the model chooses between likely candidate tokens at each prediction step – were largely unimpactful given the one-word nature of responses, but may have some small impact and are thus reported for reproducibility. Specifically, we used temperature = 0.6 (where 1.0 reproduces Thompson sampling and smaller values more heavily weight the most likely prediction), top-k = 40 (i.e., the maximum number of predicted tokens considered for sampling at each step), and top-p = .95 (i.e., the total probability mass to consider for sampling at each step).

To assess the reliability of the LLM-generated classifications, we compared the LLM predictions against human judgment on a random sample of the dataset. We used simulation-based power analysis (with 10,000 simulation iterations) to determine the sample size needed to provide a confidence interval width < 0.2 for Cohen's $\kappa$. Assuming $\kappa = .7$, this required 200 samples. We therefore manually coded 200 randomly selected instances for the four constructs. Interrater reliability showed high agreement between humans and LLM for the more objective constructs "explicit question" ($\kappa = .778$) and "discussing math" ($\kappa = .722$). There was, however, more moderate agreement for "need help" ($\kappa = .562$) and "might be confused" ($\kappa = .531$).

### 2.4.3. Linear statistical models

We measured differences between time periods with linear mixed effects models to determine whether there were significant differences in the audio, video, and transcript data in response to teacher–group interactions. Specifically, for each aggregated variable from each modality, we fit a linear model with the variable as the outcome and a dummy-coded predictor of time period (either before vs. during or before vs. after). We included a random intercept for each group recording ID to account for statistical dependence within groups, such as the possibility that the audio volume may have been higher for some recordings relative to others. We used the *lmerTest* package (version 3.1.3) in R version 4.5.1 (Kuznetsova et al., 2017; R Core Team, 2020).

The sample sizes for tests were relatively large, so sampling distributions were roughly normal per the central limit theorem, but still detectably different from normal via Shapiro–Wilk tests in most cases. We thus also repeated tests with the *rankit* transformation (Soloman & Sawilowsky, 2009), which revealed additional significant differences for some keypoint variables. However, we focused on the original, untransformed data in this paper, as the interpretation does not differ, and these are the more conservative results.

## 2.5. Qualitative Analysis of Computationally Determined Clips

We conducted small-scale qualitative analysis on a subset of data to richly describe what might be happening in the computationally detected clips and facilitate human interpretation of the clips and computational output. We used a combination of coding-oriented and insight-oriented qualitative approaches (Scherr, 2009). For both types of analysis, we used a subset of the data with an interaction clip as the primary unit of analysis. We randomly selected an interaction clip from each of the 21 computationally determined interaction clip outputs, and subsequently, before, during, and after video clips were trimmed from the original videos (resulting in 63 video segments). An overlay of the same boundary box from the pose detection process was drawn around the target group to focus on the target group, while also allowing the researcher to maintain awareness of the surrounding events outside of that boundary.

The qualitative coding focused on characterizing features of the teacher-student interaction that had been considered in the design of the detector: *teacher presence*, *student gaze*, and *overt help-seeking*. One researcher watched and coded each of these 63 video segments, applying the codes from the codebook in Table 2. As our goal was primarily descriptive rather than predictive, we did not have multiple coders to explore inter-coder agreement.

In addition to the qualitative coding, we conducted descriptive qualitative analysis of each of the 21 interaction clips. This analysis aimed to identify themes in the types of teacher-student activity, primarily focusing on the 21 during clips and use the before and after segments as context. In particular, we explored themes in relation to the aim of the computational detection, which included both similarities and contrasts with that aim.

Table 2: Codebook with behavioral constructs and definitions to assess video clips.

| Behavior | Definition | Codes |
|---|---|---|
| Gaze | A majority (i.e., two if group of three, three if group of four) or more students in the group are exhibiting the gaze characteristic for a majority of the duration (over 50%) | at task<br>at group-member<br>at teacher<br>away-distracted<br>other (if unclear) |
| Overt help-seeking | Student tries to get attention of teacher by raising hand, clearly calling for teacher | 1 for yes<br>0 for no |
| Teacher presence | Teacher is present within the boundary box, even briefly | 1 for yes<br>0 for no |

## 3. Results

In this section, we present the results of our case study. We describe the resulting clips data, including descriptive statistics about their nature, present findings on the inferential analysis from the linear mixed effects models, and present the qualitative assessment coding results.

## 3.1. Computationally Determined Clips Outputs

Applying our computationally determined clips method resulted in the identification of a total of 317 distinct teacher–group interaction events across the 21 student groups. The number of interactions per group were highly variable, with an average of 15.10 events (*SD*=12.42). One group had as few as a single event, while another had as many as 46. Similarly, the mean duration of each interaction event was 32.73 seconds (*SD*=29.35), with clips having a broad range from 10.08 to 230.36 seconds. Keeping in mind the mean duration of the analyzed videos of ~90 minutes (precisely, 1 hour, 29 minutes, and 47 seconds), the mean start time for all computationally determined clips was 52 minutes and 16 seconds (means for Teacher 1=50 min 57 sec, Teacher 2=53 min 7 sec, and Teacher 3=51 min 29 sec).

To provide an overview of the data across the different modalities, Table 3 presents the descriptive statistics for selected audio and transcript-based features for each time period across before, during (i.e., computationally determined interaction segments), and after. Among eGeMAPS' 23 low-level descriptors, in the table we present a few select features that are readily interpretable. To further clarify a few of these measures, jitter refers to vocal stability (microvariations in pitch), shimmer is vocal quality (variation in loudness), and MFCCs, or the Mel-Frequency Cepstral Coefficients, are the energy shape of the sound (how muffled or bright it sounds). We also present the mean values for four LLM-based binary outputs, which were dialogues around question-asking, detecting confusion, needing help, and math-related talk.

Table 3: A table of means for acoustic information and transcript features across the three segments of "before", "during", and "after".

| Modality | Selected features | Mean "before" | Mean "during" | Mean "after" |
|---|---|---|---|---|
| Acoustic / prosodic<br>*Based on openSMILE eGeMAPS* | Loudness (Loudness_sma3) | 0.33 | 0.34 | 0.34 |
| | Pitch (F0semitoneFrom27.5Hz) | 14.55 | 14.97 | 14.6 |
| | Jitter (jitterLocal_sma3nz) | 0.03 | 0.03 | 0.03 |
| | Shimmer (shimmerLocaldB_sma3nz) | 0.656 | 0.670 | 0.66 |
| | MFCC 1 (mfcc1_sma3) | 21.20 | 21.41 | 21.11 |
| Transcript<br>*Based on LLM outputs* | Includes question | 0.505 | 0.552 | 0.543 |
| | Dialogue indicating confusion | 0.865 | 0.861 | 0.804 |
| | Dialogue indicating needs help | 0.621 | 0.653 | 0.588 |
| | Dialogue indicating math-related talk | 0.635 | 0.653 | 0.608 |

We found no statistically significant differences comparing the *before vs. during* and *before vs. after* across all audio and transcript-based features. Paired-samples t-tests compared the selected features across the different time periods. Specifically, in the *before vs. during* comparison, none of the changes in the eGeMAPS audio features were significant, including: loudness ($t$(618) = -0.52, $p$ = .601), pitch ($t$(618) = -0.82, $p$ = .415), jitter ($t$(618) = -0.76, $p$ = .446), shimmer ($t$(618) = -0.60, $p$ = .550), MFCC 1 ($t$(618) = -0.94, $p$=.347). Similarly, there were no changes in the transcript-based features: includes question ($t$(632) = -1.19, $p$ =

.233), confused text (*t*(538) = 0.13, *p* = .898), needs help text (*t*(538) = -0.72, *p* = .473), and math talk text (*t*(538) = -0.90, *p* = .368).

Similarly, the *before vs. after* comparison showed no significant changes for any of the same features: loudness (*t*(618) = -0.35, *p* = .725), pitch (*t*(618) = -0.14, *p* = .886), jitter (*t*(618) = 0.01, *p* = .995), shimmer (*t*(618) = -0.05, *p* = .960), MFCC 1 (*t*(618) = 0.38, *p* = .703), includes question (*t*(632) = -0.95, *p* = .341), confused text (*t*(538) = 1.51, *p* = .132), needs help text (*t*(538) = 0.44, *p* = .660), and math talk text (*t*(538) = 0.53, *p* = .596).

### 3.2. Detector Validation and Coding Gazes in Interaction Video Clips

To assess the reliability and construct validity of our computer-led exploration stage, we conducted two evaluations: (1) a ground truth validation to estimate recall (i.e., how many true pedagogical interactions were found), and (2) a qualitative coding of detected clips to estimate precision and characterize student behavior. Across the three teachers' videos, the human coder identified 20 distinct interactions for the chosen group, and the detector identified 11 (55%) of these events. We analyzed the 9 missed interactions, which revealed that the detector's performance was heavily dependent on visual constraints. The method performed well when there was clear visibility (e.g., detecting 5 out of 7 events for Teacher 2) but systematically filtered out specific types of interactions. Three events were not detected as a result of the spatial boundaries of the group's pose data, meaning that the teacher stood outside of the defined group boundary box while interacting (e.g., calling on the group from the front of the class, discussing with the group while stationed at another table). Four events were missed because the teacher was either visually obscured by students/tables, or frequently crouching, behaviors that were not explicitly accounted for in the detector's heuristic for verticality. Two events were missed without a clear visual cause.

To assess the quality of the events which *were* detected, we qualitatively coded (see Table 2) a random sample of 63 video segments (21 interaction events x 3: before, during, after). In these segments, students most often looked at their materials: 30 of 63 segments (47.6%) were labelled "at task" gaze, "at teacher" in 16 segments (25.4%), "at peer" in 10 segments (15.9%), "away-distracted" in four segments (6.3%), and ambiguous "other" gazes in three segments (4.8%). Overt help-seeking was virtually nonexistent, appearing in just one of 63 segments (1.6%), and the teacher was present within the group boundary in 31 segments (49.2%).

### 3.3. Descriptive Qualitative Findings

From the insight-oriented qualitative analysis, we highlight two distinct characterizations of the teacher and student activity in the episodes that contrast with one another: joint attention to a piece of student work and parallel presence of the teacher and students. We present short examples of each, highlighting the contrast in each, and use the qualitative codes of teacher presence, overt help-seeking, and gaze as an analytic framework to describe them. Common to both types, the teacher was physically present in the group boundary and in physical proximity to the students. Another similarity was that neither type involved overt help-seeking from students.

A key difference between the types was highlighted by gaze. In the joint attention episodes, the teacher and students had their gaze fixed on something in common (often aligning with the "at task" qualitative code), which was interspersed with gaze directed at each other. In one example, this was a piece of written work a student had been working on at their table that the teacher and students discussed and referenced. In a different example, the students and the teacher were discussing work on the board in the classroom. In contrast, the parallel presence episodes did not feature joint attention, and the teacher's (and sometimes students') gaze was not focused on each other. In one example, while a student was presenting at the front of the room, the teacher sat down with a small group at the back of the classroom, occasionally facilitating the whole class discussion of the student's presentation. Thus, the teacher's and students' gazes in the group were directed at the front of the room toward the presenting student and other students in the classroom, and the teacher's comments were not directed specifically at the students in the group she was sitting with. We interpret these findings as presenting the possibility that the clip detector may be better described as detecting the co-presence of the teacher and students. The new aim and description could lead to future explorations of features such as teacher-student proximity (Heng et al., 2017) and classroom geographies (Shapiro et al., 2024) as related to teacher-student group interactions.

## 3.4. Inferential Analysis

The mixed-effects models revealed several statistically significant changes, with most of the effects occurring around the students' pose, and a specific change in the LLM-generated confused feature. Table 4 summarizes the full results for significant features across the two contrasts of interest: *before vs. during* and *before vs. after*. The analysis revealed significant shifts in both the vertical and horizontal positions of students' heads and a few body keypoints.

Table 4: Significant features ($\alpha$ = .05) from the mixed effects analysis.

| **Comparison** | **Features**<br>(keypoint num in parenthesis) | **Coefficient** ($\beta$) | ***t*-value** | ***p*-value** |
|---|---|---|---|---|
| *before vs. during* | Nose (0) x | 13.64 | 2.67 | .008 |
| | Nose (0) y | 11.13 | 3.86 | <.001 |
| | Left elbow (6) y | 6.225 | 2.03 | .042 |
| | Left wrist (7) y | 8.942 | 2.00 | .046 |
| | Right eye (15) x | 16.16 | 2.80 | .005 |
| | Right eye (15) y | 12.17 | 3.89 | <.001 |
| | Left eye (16) x | 20.79 | 3.25 | .001 |
| | Left eye (16) y | 16.56 | 4.48 | <.001 |
| | Right ear (17) x | 14.10 | 2.18 | .030 |
| | Right ear (17) y | 10.56 | 2.76 | .006 |
| | Left ear (18) x | -14.82 | -2.26 | .024 |
| *before vs. after* | Nose (0) x | 14.53 | 2.72 | .006 |
| | Nose (0) y | 12.57 | 4.17 | <.001 |
| | Left eye (16) x | 16.64 | 2.61 | .009 |
| | Left eye (16) y | 13.51 | 3.56 | <.001 |
| | Left small toe (20) y | 10.06 | 1.97 | .049 |
| | Dialogue indicating confusion | -0.061 | -1.96 | .049 |

In the *before vs. during* comparison, 11 keypoint coordinates changed significantly, and five keypoint coordinates for the *before vs. after* comparison. No statistically significant changes were found for any of the audio features in either comparison. For the transcript-based features, a single significant effect was found: the confused feature showed a significant decrease in the before vs. after comparison ($\beta$=-0.061, $t$=-1.96, $p$=.049).

# 4. Discussion

In this section, we discuss the results of our case study. We first discuss a lens for considering assumptions from researchers as well as algorithms, and present our simple framework of the three stages of data analysis: exploration, interpretation, and confirmation. We then discuss our case study results within this framework, which is followed by a thought experiment to envision how the results and stages might have differed if the leading entity were interchanged during the stages.

## 4.1. Researchers' and Algorithms' Lenses

Researchers bring their own "positionality," i.e., their backgrounds, theoretical views, and cultural experiences, to any study (Milner IV, 2007). Positionality shapes how researchers define problems, select data, and interpret results. Researcher positionality is well established in qualitative work, where personal insight and context are acknowledged as part of the inquiry process (Arnould et al., 2006; Snape & Spencer, 2003). Yet, even in quantitative studies, decisions about data selection, variable construction, and model choices are not neutral. Each choice reflects assumptions about what is important and what counts as evidence. In computational research, these decisions are often embedded in algorithms, making them less visible but still influential. Algorithms therefore also carry their own kind of positionality, as they are built to detect certain types of patterns based on training data, design choices, or the mathematical structures they use – the lens through which they are able to explore, interpret, or confirm patterns in data. For example, a video analysis algorithm such as one we described in this paper to identify teacher–group interactions may miss verbal exchanges that happen across the room, where the teacher is not physically near the group.

Standard methods in educational data mining also favor specific types of patterns. For instance, clustering algorithms must, by definition, assume something about the "shape" of a cluster in data – e.g., that groups are compact and well-separated (Xu & Wunsch, 2005), which works well for some data but less well for others. An overarching fundamental problem in search (e.g., clustering, exploration) and optimization (e.g., classification, regression) is the no free lunch theorem (Wolpert & Macready, 1997). The no free lunch theorem proves that no search/optimization method is better than any other *in general*, only in specific contexts, precisely because any useful lens an algorithm (or human) brings to an analysis will equally as much lead them astray in a contrasting context. It is therefore essential to consider and transparently report the role that humans and algorithms have in each step of data analysis, given the impact that the biases common to broad classes of methods (e.g., cognitive biases for humans) may have on results.

Considering who/what leads each step of analysis is also important for understanding the transparency of an analysis, given that each step of analysis has the potential to impact transparency in later steps. For instance, a predictive model of student outcomes may be of a highly interpretable form, such as a simple decision tree, yet still result in a completely uninterpretable final decision because an earlier step of the process (i.e., data exploration for feature engineering) was led by a mostly uninterpretable algorithm (Tang & Bosch, 2024; Bosch, 2021). Understanding the role of both human and algorithmic agency in each stage of analysis, therefore, helps researchers make informed choices about method design and interpretation, especially when comparing findings across different studies where differing defaults or assumptions may shape results.

As summarized in Section 1.1, our analytic agency framework makes these lenses explicit by treating exploration, interpretation, and confirmation as stages that can be primarily human-led or computer-led, and we situate common methods and our own case study within that framework (Table 1). Below, we use that framework to understand the context of our findings and to explore counterfactual combinations of stage leadership.

## 4.2. Our Results Through The Lens of Exploration, Interpretation, Confirmation

Viewed through this framework, our results are consistent with an analytic workflow in which exploration (e.g., feature extraction, interaction determination) and much of interpretation (e.g., statistical analyses) were effectively *computer-led*. Our teacher–group interaction detector (presented in detail in Section 2.3) presented 317 teacher–group interaction events across the 21 groups. These distributions reflect the design of our detection method: interaction candidates were defined by pose-based vertical dominance and proximity, plus a 10s smoothing window and 10s minimum episode length. In other words, the *tool*, rather than the researcher, set the decision boundary for what counted as an "interaction." Similarly, it could be argued that some of the hard decision boundaries were also influenced by the computer, through the sheer process of pose feature extraction on the video using OpenPose. Since OpenPose was trained on large-scale datasets (i.e., COCO, Lin et al., 2014; and MPII, Andriluka et al., 2014) with predefined keypoints and thresholds, it outputs inherent assumptions about what bodily configurations are detectable and common, which likely further shaped which events surfaced in our pipeline.

Where change appeared, as naturally expected, it mostly aligned with the modality the detector used as a foundation. Although comparing *before vs. during* and *before vs. after* did not yield significant effects, mixed-effects models showed significant shifts based on the specific body keypoints that we had privileged in our detector design. Many of the keypoints that we had used for the head centroid (nose, neck, right eye, left eye, right ear, left ear) were found in the list of significant keypoints in Table 4. This pattern is what our pose-oriented exploration lens makes visible: body and spatial aspects as the teacher approaches and departs. Interestingly, the LLM-generated confused text feature was also found to be significant ($\beta = 0.061$, $t = 1.96$, $p = .049$) for *before vs. after*, meaning that the presence of confusion in student conversations decreased after teacher–group interactions, on average, suggesting that teacher intervention in student groups helped them to overcome impasses they were encountering and move on.

By contrast, the modalities our pipeline represented more coarsely did not show significant effects. eGeMAPS features were essentially flat across segments. Taken together, the effects show where the computer led both the definition of events and the representation of meaning. Early design choices for the detector, such as the pose heuristic and smoothing thresholds may have created strong biases toward valuing spatial dynamics. This influenced the exploration process, which may have unintentionally overlooked other meanings in the data.

While our analysis did not include a formal third stage of confirmation, the light qualitative coding carried out by three researchers can be viewed as a form of human-led appraisal of the computer-generated output. Crucially, the teacher was coded as present within the group's boundary box in approximately half of the randomly selected segments (49.2%), meaning that, despite the reasonable simplicity of our head centroid heuristic, the clips trended toward a pattern of being useful for detecting the teacher in some capacity. Yet, "at teacher" gaze codes were at a lower frequency at 25.4%, suggesting our computer-led exploration frequently identified moments of mere co-presence rather than sustained interactions. Students were most often coded as looking "at task" (47.6%), instead. In essence, this human check hints that our computer-led pipeline successfully found what it was designed to find – patterns in spatial data – but that these patterns are an incomplete and sometimes misleading proxy for the richer construct of teacher–student interaction.

## 4.3. Counterfactuals for Analysis Leadership

To make the impact of analytical agency concrete, we can conduct a thought experiment, exploring counterfactual research designs for our case study where the leading entity (i.e., human or computer) for exploration and interpretation is deliberately shifted. Using our same data and research goal, to detect and understand teacher–student group interactions, we can imagine how different choices of leadership would have produced fundamentally different studies and findings.

### 4.3.1. Human-led exploration, computer-led interpretation

First, let us consider a design of *human-led* exploration and *computer-led* interpretation. In this scenario, a researcher with pedagogical expertise would begin the analysis, not an algorithm. They would watch the raw classroom videos and, guided by their theoretical knowledge of responsive teaching and classroom discourse, likely manually identify and timestamp moments of what they judge to be meaningful teacher–student interaction. These human-identified events might include moments the computer missed, such as a teacher offering brief verbal feedback from across the room, or moments the computer flagged incorrectly, like a student standing up to stretch near the teacher. Even short but meaningful teacher-group interactions would be preserved, which the computational approach forwent by filtering potential interactions under 10 seconds to reduce false positives. Additionally, as humans can view interactions to collect with more nuance, clear misclassifications based on purely pose orientations would not occur, such as when a non-teacher individual (e.g., student) is standing to talk to a friend at another group's table (Figure 5, left), or a teacher is sitting down nearby a group's desk (Figure 5, right). These qualitatively selected clips could then be fed into the same computational pipeline we used: openSMILE for extracting acoustic features and an LLM for transcript text analysis.

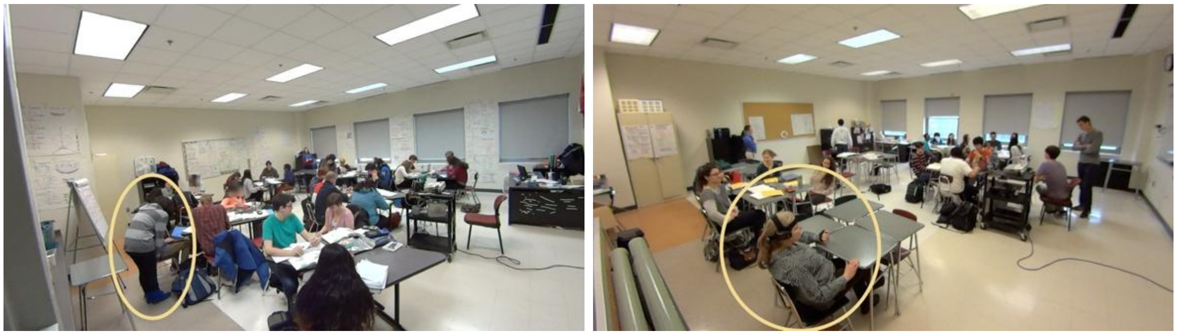


Figure 5: Examples of when a computer-led pose-based interaction detection might misclassify. An example of when a student might be standing around a group's desk, and an example of when a teacher might be sitting down nearby a group's desk.

A reasonable outcome could be a reversal of our findings. We would expect to see much stronger and more interpretable statistical signals in the audio and transcript data. Because the initial exploration was guided by human meaning-making, the subsequent computer-led interpretation would be quantifying phenomena already deemed salient by an expert. We might find, for instance, a statistically significant increase in the “Needs help text” feature *before* the teacher’s arrival, followed by a noticeable decrease *after*, which may be observed with an increase in “math talk” *during* the actual interaction. While perhaps more difficult to predict, there may also be noticeable differences in the segments for the prosodic and acoustic eGeMAPS features. For example, perhaps the speech patterns before a teacher’s interaction with the group may be characterized by more shimmer due to a student feeling uncertain about a topic, which leads to less consistent loudness in speech. It could also be the case that students are speaking with more energy after an interaction, due to newfound confidence about the topic. The computer can further interpret the human-led and identified interactions, not as a discoverer but a magnifier by adding statistical rigor to patterns first identified by the human lens. While the study’s validity would rest on the qualitative expertise of the human coder, it is possible to imagine that there would be clearer and pedagogically relevant patterns.

### 4.3.2. Computer-led exploration, human-led interpretation

Second, let us imagine the inverse: *computer-led* exploration and *human-led* interpretation. This approach begins exactly as our case study did, using the pose-based detector algorithm to generate the 317 interaction clips. However, instead of proceeding to statistical analysis, these clips would be handed to a qualitative researcher as the primary dataset for analysis. The human would take the algorithm’s output as a pre-filtered, high-volume sample of potentially interesting moments that would be infeasible to find manually.

The analytical process would then shift entirely. A researcher would perform a deep, qualitative analysis – such as conversation analysis or interaction analysis – on the computer-generated segments (Seedhouse, 2005; Norcutt & McCoy, 2004). Instead of *p*-values and coefficients, the output would be potentially thematic codes, case descriptions, and narrative accounts of the interactions. The researcher might discover that the pose-based, physical proximity interactions identified by the algorithm are disproportionately procedural (e.g., checking answers to problem sets, distributing worksheets) rather than conceptual (e.g., explaining a specific concept in detail to the group). Or they might find that even within these algorithmically similar clips, the teacher uses a variety of pedagogical moves that lead to different student responses. In this design, the computer acts as a tireless, albeit naive, research assistant to direct the limited and expensive resource of human attention. The study’s validity would depend on the depth and rigor of the human’s qualitative interpretation of the machine-selected data.

### 4.3.3. Iterative hybrid partnership

Finally, we can envision our case study as having been an *iterative hybrid partnership* of computer-led exploration, human-led interpretation, similar in vein to computational grounded theory (Nelson, 2020). We could begin with our computer-led exploration. A human researcher could then interpret this initial output, labeling the algorithm’s findings as “Yes, this was a true interaction, so 1,” or “No, this was just the teacher walking by so 0,”, but perhaps with more systematic but nuanced codes like “procedural check-in,” “conceptual explanation,” or “off-task monitoring.” This human feedback could be used to fine-tune the detection algorithm or train a machine learning model for a more sophisticated approach to predict the interactions. The refined approach could then perform a second, more intelligent exploration of the data and present a new set of clips for human interpretation. An additional direction could be that the human researcher might notice in the first set of clips that a conceptual breakthrough is preceded by certain acoustic patterns, such as lower speech energy and more shimmer in student voices, perhaps indicative of uncertainty. The refined algorithm could then be trained to search the entire dataset for this acoustic signature of this uncertainty measure, which, when followed by a teacher's presence, could lead to a more targeted and meaningful set of interactions for qualitative analysis.

Similarly, the researcher could use the transcripts to identify key conversation markers of scaffolding, such as teachers using phrases like "What if you tried...?" or "What does that number tell you?" A refined algorithm could associate these phrases with true interaction labels to find similar moments of high-quality teacher–group talk that the initial pose-only detector missed. The cycle of computer exploration and human interpretation would allow the researcher's theoretical understanding of classroom discourse to evolve with the process. This type of stage-level delegation mirrors work by Barany et al. (2024), who empirically compared four workflows for qualitative codebook development, ranging from fully human to fully automated (human only, human development–ChatGPT refinement, ChatGPT development–human refinement, ChatGPT only). The finding showed that hybrid workflows, where the agency alternated between human and AI, produced the highest quality codebooks, which supports our perspective that intentional, explicit partnerships often yield more meaningful inquiry than fully automated or fully manual alternatives. For educational data mining research, this framework provides a way to regard computational tools in research pipelines also as influencing analytic agency and the eventual results obtained. By making this analytic agency an explicit design choice, researchers could better align computational methods with theoretical aims, leading to more transparent findings in classroom environments.

These counterfactuals demonstrate that the choice of who or what leads each analytical stage is a fundamental decision point that defines the nature of the research. Our framework does not prescribe one path as superior, but instead provides the language for researchers to navigate these choices deliberately. Through the perspective provided by the framework, we can perhaps better anticipate and see the inherent biases and positionalities of the researcher, and how their methodological partnership with computational tools and AI lead them to the eventual results.

## 5. Limitations and Conclusion

In this paper, we presented a study that automatically detected teacher–group interactions from high school mathematics classroom videos. We developed and applied a computer-led pipeline to detect teacher-student group interactions based on pose data extracted from the video data. Our case study illustrates a simple framework for considering researcher analytical agency during the exploration, interpretation, and confirmation stages of data analysis. The results were consistent with the principles of our framework: the exploration stage, which was driven by a pose-based algorithm, yielded statistically significant findings primarily in the pose data itself, while showing no significant effects in audio, and a significant effect in only one transcript-based feature, dialogue about confusion. The lead entity for exploration and interpretation, in our case, the computer, shaped the patterns that were discovered and considered meaningful. This study contributes to the EDM field by providing the vocabulary needed to describe these shifts in agency and the inherent assumptions of multimodal computational tools. By classifying whether each stage is primarily human-led or computer-led, our framework provides a vocabulary needed to describe and compare the underlying assumptions of different research methods when integrating computational tools and AI during the data analysis process.

We acknowledge several limitations in our study. First, our interaction detector used a simple pose-based heuristic (vertical dominance and proximity), which we view as a coarse proxy for teacher–group co-presence rather than a validated interaction. As our own qualitative coding suggested, this simplicity contributed to both false positives (e.g., brief walk-bys, procedural check-ins) and missed interactions (e.g., seated or across-room exchanges). Yet, this limited construct validity illustrated how computer-led exploration can narrow the space of events that are available for later interpretation. As an example of a way to extend on this detector, rather than using just pose, the approach could have integrated transcripts and even acoustic characteristics, to be used in conjunction to identify clearer signals for teacher interactions. Furthermore, our statistical analysis did not include corrections for multiple comparisons, as its primary purpose was to demonstrate the assumptions in the pose-based approach. This does imply that some of the detailed findings may not be reliable, and the overall patterns and trends should be treated as the primary contribution of this work rather than treating every single statistical finding as reliable. Finally, our three-stage framework is a simple model; real-world research is often a more complex and iterative process that may not fit neatly. Yet, the simplicity of the framework likely makes analytical leadership more transparent and interpretable across diverse methods.

Future research should empirically investigate the counterfactual scenarios outlined in our discussion, for instance, by directly comparing the findings of a specific study from a human-led exploration of the same data against computer-led results. Further work could also focus on developing tools that are designed to support hybrid, iterative research partnerships between human and computational tools. These tools should crucially be able to document the decisions made by the researcher and computational tools to make positionality and inherent assumptions clearer. Ultimately, by making analytical agency an explicit design choice, researchers can combine the complementary strengths of human insight and computational power, and work toward constructive and transparent ways to derive meaningful insights from the data at hand.

## 6. Acknowledgements

This material is based upon work supported by the National Science Foundation (DRL-1920796). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

Ahuja, K., Kim, D., Xhakaj, F., Varga, V., Xie, A., Zhang, S., Townsend, J. E., Harrison, C., Ogan, A., and Agarwal, Y. 2019. EduSense: Practical classroom sensing at scale. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(3), 1–26. https://doi.org/10.1145/3351229

Alderete, J., Hui, M. K. F., and Mohan, A. 2025. *Evaluating ASR robustness to spontaneous speech errors: A study of WhisperX using a speech error database*. arXiv. https://doi.org/10.48550/arXiv.2508.13060

Alibali, M. W., and Nathan, M. J. 2012. Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences*, *21*(2), 247–286. https://doi.org/10.1080/10508406.2011.611446

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. 2014. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE*

*Conference on Computer Vision and Pattern Recognition*, 3686–3693. https://doi.org/10.1109/CVPR.2014.471

ARNOULD, E., PRICE, L., AND MOISIO, R. 2006. Making contexts matter: Selecting research contexts for theoretical insights. In *Handbook of Qualitative Research Methods in Marketing*, R. W. Belk, Ed. Edward Elgar Publishing. 106–128. https://doi.org/10.4337/9781847204127.00016

BAIN, M., HUH, J., HAN, T., AND ZISSERMAN, A. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of INTERSPEECH 2023*, 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

BAKER, R. S., HUTT, S., BROOKS, C. A., SRIVASTAVA, N., AND MILLS, C. 2024. Open science and educational data mining: Which practices matter most? In *Proceedings of the 17th International Conference on Educational Data Mining*, C. Demmans Epp, B. Paaßen, and D. Joyner, Eds. International Educational Data Mining Society, 279–287. https://doi.org/10.5281/zenodo.12729816

BARANY, A., NASIAR, N., PORTER, C., ZAMBRANO, A. F., ANDRES, A. L., BRIGHT, D., SHAH, M., LIU, X., GAO, S., ZHANG, J., MEHTA, S., CHOI, J., GIORDANO, C., AND BAKER, R. S. 2024. ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In *Artificial Intelligence in Education. AIED 2024* (Lecture Notes in Computer Science, Vol. 14830), A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds. Springer, Cham, Switzerland, 134–149. https://doi.org/10.1007/978-3-031-64299-9_10

BERGNER, Y., GRAY, G., AND LANG, C. 2018. What does methodology mean for learning analytics? *Journal of Learning Analytics*, *5*(2), 1–8. https://doi.org/10.18608/jla.2018.52.1

BOERSMA, P., AND VAN HEUVEN, V. 2001. Speak and unSpeak with Praat. *Glot International*, *5*(9/10), 341–347.

BOSCH, N. 2021. AutoML feature engineering for student modeling yields high accuracy, but limited interpretability. *Journal of Educational Data Mining, 13*(2), 55–79. https://doi.org/10.5281/zenodo.5275314

BREDIN, H., YIN, R., CORIA, J. M., GELLY, G., KORSHUNOV, P., LAVECHIN, M., FUSTES, D., TITEUX, H., BOUAZIZ, W., AND GILL, M. P. 2020. pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

CAO, Z., HIDALGO, G., SIMON, T., WEI, S. E., AND SHEIKH, Y. 2021. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

CHARMAZ, K. 2014. *Constructing Grounded Theory* (2nd ed.). SAGE Publications.

CHOI, Y., LEE, Y., SHIN, D., CHO, J., PARK, S., LEE, S., BAEK, J., BAE, C., KIM, B., AND HEO, J. 2020. EdNet: A large-scale hierarchical dataset in education. In *Artificial Intelligence in Education. AIED 2020* (Lecture Notes in Computer Science, Vol. 12164), I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Springer, Springer, Cham, 69–73. https://doi.org/10.1007/978-3-030-52240-7_13

COHN, C., DAVALOS, E., VATRAL, C., FONTELES, J. H., WANG, H. D., MA, M., AND BISWAS, G. 2024. *Multimodal methods for analyzing learning and training environments: A systematic literature review*. arXiv. https://doi.org/10.48550/arXiv.2408.14491

D'MELLO, S. K., AND GRAESSER, A. 2023. Intelligent tutoring systems: How computers achieve learning gains that rival human tutors. In *Handbook of Educational Psychology* (4th ed.), P. A. Schutz and K. R. Muis, Eds. Routledge, New York, 603–629. http://doi.org/10.4324/9780429433726-31

DRAGUT, E., LI, Y., POPA, L., AND VUCETIC, S. 2021. Data science with human in the loop. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 4123–4124. https://doi.org/10.1145/3447548.3469476

EYBEN, F., SCHERER, K. R., SCHULLER, B. W., SUNDBERG, J., ANDRÉ, E., BUSSO, C., DEVILLERS, L. Y., EPPS, J., LAUKKA, P., NARAYANAN, S. S., AND TRUONG, K. P. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

EYBEN, F., WÖLLMER, M., AND SCHULLER, B. 2010. openSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. https://doi.org/10.1145/1873951.1874246

FRANKEL, L., AND BROWNSTEIN, B. 2016. An evaluation of the usefulness of prosodic and lexical cues for understanding synthesized speech of mathematics. *ETS Research Report Series*, *2016*(2), 1–19. https://doi.org/10.1002/ets2.12119

FRENCH, D., MOULDER, R., EZEMA, K., VON DER WENSE, K., AND D'MELLO, S. 2025. Linguistic alignment predicts learning in small group tutoring sessions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Association for Computational Linguistics, Suzhou, China, 15600–15611. https://doi.org/10.18653/v1/2025.findings-emnlp.844

GABBAY, H., AND COHEN, A. 2022. Investigating the effect of automated feedback on learning behavior in MOOCs for programming. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society, 376–383. https://doi.org/10.5281/zenodo.6853124

GONZALES, A. C., PURINGTON, S., ROBINSON, J., AND NIESWANDT, M. 2019. Teacher interactions and effects on group triple problem solving space. *International Journal of Science Education, 41*(13), 1744–1763. https://doi.org/10.1080/09500693.2019.1638982

GONZÁLEZ-BRENES, J. P., AND MOSTOW, J. 2012. Dynamic cognitive tracing: Towards unified discovery of student and cognitive models. In *Proceedings of the 5th International Conference on Educational Data Mining*, K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, and J. Stamper, Eds. International Educational Data Mining Society, 49–56.

HAIDER, F., POLLAK, S., ALBERT, P., AND LUZ, S. 2021. Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech & Language*, *65*, Article 101119. https://doi.org/10.1016/j.csl.2020.101119

HENG, B. C., CHEONG, C. Y. M., AND TAIB, F. 2017. Instructional proxemics and its impact on classroom teaching and learning. *International Journal of Modern Languages and Applied Linguistics*, *1*(1), 69–85. https://journal.uitm.edu.my/ojs/index.php/IJMAL

HUR, P., AND BOSCH, N. 2022. Tracking individuals in classroom videos via post-processing OpenPose data. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, 465–471. https://doi.org/10.1145/3506860.3506888

KAENDLER, C., WIEDMANN, M., RUMMEL, N., AND SPADA, H. 2015. Teacher competencies for the implementation of collaborative learning in the classroom: A framework and research review. *Educational Psychology Review*, *27*(3), 505–536. https://doi.org/10.1007/s10648-014-9288-9

KOCABAS, M., ATHANASIOU, N., AND BLACK, M. J. 2020. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5253–5263. https://doi.org/10.1109/CVPR42600.2020.00530

KUMAR, D., MADAN, S., SINGH, P., DHALL, A., AND RAMAN, B. 2024. Towards engagement prediction: A cross-modality dual-pipeline approach using visual and audio features. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11383–11389. https://doi.org/10.1145/3664647.3688986

KUZNETSOVA, A., BROCKHOFF, P. B., AND CHRISTENSEN, R. H. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

LIN, R., AND KOEDINGER, K. R. 2017. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining*, *9*(1), 25–41. https://doi.org/10.5281/zenodo.3554625

LIN, T. Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014* (Lecture Notes in Computer Science, Vol. 8693) D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, Cham, Switzerland, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

LU, W., LAFFEY, J., SADLER, T., GRIFFIN, J., AND GOGGINS, S. 2024. A scalable, flexible, and interpretable analytic pipeline for stealth assessment in complex digital game-based learning environments: Towards generalizability. *Journal of Educational Data Mining*, *16*(2), 149–176. https://doi.org/10.5281/zenodo.14503598

MEJIA-DOMENZAIN, P., NAZARETSKY, T., SCHULTZE, S., HOCHWEBER, J., AND KÄSER, T. 2024. Navigating self-regulated learning dimensions: Exploring interactions across modalities. In *Artificial Intelligence in Education. AIED 2024* (Lecture Notes in Computer Science, Vol. 14830), A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds. Springer, Cham, Swizerland, 104–118. https://doi.org/10.1007/978-3-031-64299-9_8

MILLS, C., GREGG, J., BIXLER, R., AND D'MELLO, S. K. 2021. Eye-Mind Reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction*, *36*(4), 306–332. https://doi.org/10.1080/07370024.2020.1716762

MILNER IV, H. R. 2007. Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher*, *36*(7), 388–400. https://doi.org/10.3102/0013189X07309471

MISTRAL AI. 2025. *Mistral Small 3.2 24B Instruct 2506 [Large language model].* Hugging Face. https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

NELSON, L. K. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

NORCUTT, N., AND MCCOY, D. 2004. *Interactive Qualitative Analysis: A Systems Method for Qualitative Research*. SAGE Publications. https://doi.org/10.4135/9781412984539

OCHOA, X., AND WORSLEY, M. 2016. Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, *3*(2), 213–219. https://doi.org/10.18608/jla.2016.32.10

OUHAICHI, H., BAHTIJAR, V., AND SPIKOL, D. 2024. Exploring design considerations for multimodal learning analytics systems: An interview study. *Frontiers in Education*, *9*, Article 1356537. https://doi.org/10.3389/feduc.2024.1356537

PARR, E. D. 2021. Making space for joint exploration: The embodiment of social and epistemic positioning in student-teacher interaction. In *Proceedings of the 15th International Conference of the Learning Sciences - ICLS 2021*, E. de Vries, Y. Hod, and J. Ahn, Eds. International Society of the Learning Sciences, 843–850. https://par.nsf.gov/biblio/10291504

R CORE TEAM 2020. *R: A language and environment for statistical computing* (Version 4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

RADFORD, A., KIM, J. W., XU, T., BROCKMAN, G., MCLEAVEY, C., AND SUTSKEVER, I. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*¸ A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds. PMLR, 28492–28518. http://doi.org/10.48550/arXiv.2212.04356

RAJARATHINAM, R. J., PALAGUACHI, C., AND KANG, J. 2025. 360-degree cameras vs traditional cameras in multimodal learning analytics: Comparative study of facial recognition and pose estimation. *Journal of Educational Data Mining*, *17*(1), 157–182. https://doi.org/10.5281/zenodo.14966499

ROMERO, C., AND VENTURA, S. 2020. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), Article e1355. https://doi.org/10.1002/widm.1355

SALVI, R. C., AND BOSCH, N. 2025. Investigating perception of gender stereotypes in large language models: A computational grounded theory approach. *ACM Journal on Responsible Computing*, *2*(2), 1–29. https://doi.org/10.1145/3737882

SANDVIG, C. 2014. Seeing the sort: The aesthetic and industrial defense of "the algorithm." *Media-N*, *10*(1). http://median.newmediacaucus.org/art-infrastructures-information/seeing-the-sort-the-aesthetic-and-industrial-defense-of-the-algorithm/

SCHERR, R. E. 2009. Video analysis for insight and coding: Examples from tutorials in introductory physics. *Physical Review Special Topics - Physics Education Research*, *5*(2), 020106. https://doi.org/10.1103/PhysRevSTPER.5.020106

SEEDHOUSE, P. 2005. Conversation analysis and language learning. *Language Teaching*, *38*(4), 165–187. https://doi.org/10.1017/S026144480500318X

SHAPIRO, B. R., HORN, I. S., GILLIAM, S., AND GARNER, B. 2024. Situating teacher movement, space, and relationships to pedagogy: A visual method and framework. *Educational Researcher*, *53*(6), 335–347. https://doi.org/10.3102/0013189X241238698

SHUTE, V. J. 2011. Stealth assessment in computer-based games to support learning. In *Computer Games and Instruction*, S. Tobias and J. D. Fletcher, Eds. Information Age Publishing, Charlotte, NC, 503–524.

SINGH, S., SINGH, L., AND SATSANGEE, N. 2025. Automated assessment of classroom interaction based on verbal dynamics: A deep learning approach. *SN Computer Science, 6*(3), Article 201. https://doi.org/10.1007/s42979-025-03770-3

SIVAKUMARAN, N., YANG, C. Y., ZALA, A., YU, S., HONG, D., ZOU, X., STENGEL-ESKIN, E., CARPENTER, D., MIN, W., HMELO-SILVER, C., ROWE, J., LESTER, J., AND BANSAL, M. 2025. A multimodal classroom video question-answering framework for automated understanding of collaborative learning. In *Proceedings of the 27th International Conference on Multimodal Interaction*, 516–525. Association for Computing Machinery. https://doi.org/10.1145/3716553.3750795

SNAPE, D., AND SPENCER, L. 2003. The foundations of qualitative research. In *Qualitative Research Practice: A Guide for Social Science Students and Researchers*, J. Ritchie and J. Lewis Eds. SAGE Publications, 1–23..

SOLOMAN, S., AND SAWILOWSKY, S. 2009. Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods, 8*(2), 448–462. https://doi.org/10.22237/jmasm/1257034080

TANG, L., AND BOSCH, N. 2024. Can students understand AI decisions based on variables extracted via AutoML? In *Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3342–3349. https://doi.org/10.1109/SMC54092.2024.10831034

VENKATESHA, V., BRADFORD, M., AND BLANCHARD, N. 2025. Dude, where's my utterance? Evaluating the effects of automatic segmentation and transcription on CPS detection. In *Artificial Intelligence in Education. AIED 2025* (Communications in Computer and Information Science, Vol. 2592), A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, and S. Isotani Eds. Springer, Cham, Switzerland, 144–151. https://doi.org/10.1007/978-3-031-99267-4_18

VIEIRA, F., CECHINEL, C., RAMOS, V., RIQUELME, F., NOEL, R., VILLARROEL, R., CORNIDE-REYES, H. AND MUNOZ, R. 2021. A learning analytics framework to analyze corporal postures in students' presentations. *Sensors*, *21*(4), Article 1525. https://doi.org/10.3390/s21041525

WENSKOVITCH, J., AND NORTH, C. 2020. Interactive artificial intelligence: Designing for the "two black boxes" problem. *Computer*, *53*(8), 29–39. https://doi.org/10.1109/MC.2020.2996416

WHITEHILL, J., AND LOCASALE-CROUCH, J. 2023. Automated evaluation of classroom instructional support with LLMs and BoWs: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*, *16*(1), 34–60. https://doi.org/10.5281/zenodo.10974824

WOLPERT, D. H., AND MACREADY, W. G. 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

XUE, W., CUCCHIARINI, C., VAN HOUT, R. W. N. M., AND STRIK, H. 2019. Acoustic correlates of speech intelligibility: The usability of the eGeMAPS feature set for atypical speech.

In *Proceedings of the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 48–52. https://doi.org/10.21437/SLaTE.2019-9

XU, R., AND WUNSCH, D. 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

YIN, S., LIU, Z., GOH, D. L., QUEK, C., AND CHEN, N. 2025. Scaling up collaborative dialogue analysis: An AI-driven approach to understanding dialogue patterns in computational thinking education. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 47–57. Association for Computing Machinery. https://doi.org/10.1145/3706468.3706474

YOON, S. A., AND HMELO-SILVER, C. E. 2017. What do learning scientists do? A survey of the ISLS membership. *Journal of the Learning Sciences*, *26*(2), 167–183. https://doi.org/10.1080/10508406.2017.1279546

ZHAO, J., LI, J., AND JIA, J. 2021. A study on posture-based teacher-student behavioral engagement pattern. *Sustainable Cities and Society*, *67*, Article 102749. https://doi.org/10.1016/j.scs.2021.102749