# “Beautiful work, you’re rock stars!”: Teacher Analytics to Uncover Discourse that Supports or Undermines Student Motivation, Identity, and Belonging in Classrooms


Nicholas C. Hunkins
University of Colorado Boulder
nicholas.hunkins@colorado.edu

Sean Kelly
University of Pittsburgh
spkelly@pitt.edu

Sidney K. D’Mello
University of Colorado Boulder
sidney.dmello@colorado.edu



## ABSTRACT

From carefully crafted messages to flippant remarks, warm expressions to unfriendly grunts, teachers’ behaviors set the tone, expectations, and attitudes of the classroom. Thus, it is prudent to identify the ways in which teachers foster motivation, positive identity, and a strong sense of belonging through inclusive messaging and other interactions. We leveraged a new coding of teacher supportive discourse in 156 video clips from 73 6$^{th}$ to 8$^{th}$ grade math teachers from the archival Measures of Effective Teaching (MET) project. We trained Random Forest classifiers using verbal (words used) and paraverbal (acoustic-prosodic cues, e.g., speech rate) features to detect seven features of teacher discourse (e.g., public admonishment, autonomy supportive messages) from transcripts and audio, respectively. While both modalities performed over chance guessing, the specific language content was more predictive than paraverbal cues (mean correlation = .546 vs. .276); combining the two yielded no improvement. We examined the most predictive cues in order to gain a deeper understanding of the underlying messages in teacher talk. We discuss implications of our work for teacher analytics tools that aim to provide educators and researchers with insight into supportive discourse.



## CCS CONCEPTS

• **Applied Computing**; • **Education**; • **Computing Methodologies**; • **Machine Learning**;

## KEYWORDS

Teacher Analytics, Discourse Analysis, Natural Language Processing, Automated Feedback

**ACM Reference Format:**

Nicholas C. Hunkins, Sean Kelly, and Sidney K. D’Mello. 2022. “Beautiful work, you’re rock stars!”: Teacher Analytics to Uncover Discourse that Supports or Undermines Student Motivation, Identity, and Belonging in Classrooms. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3506860.3506896





*LAK22, March 21–25, 2022, Online, USA*



ACM ISBN 978-1-4503-9573-1/22/03.
https://doi.org/10.1145/3506860.3506896


## 1 INTRODUCTION

The teacher-student relationship, ever present across the myriad classroom interactions, is crucial for effective teaching and learning. Considerable research shows that effective teaching requires support for students’ socio-emotional needs as well as their academic achievement [1, 2, 7, 20, 28, 35]. Students’ sense of belonging is related to a range of outcomes, including academic persistence, mental health, and educational aspiration [27, 49], but students who feel they are not cared for in school are at risk of feeling alienated or isolated. For example, “Is this a safe learning space *for me* or do I not belong here?” is a question often asked by students, especially those with stigmatized identities (e.g., girls in math classes [43]). Alongside structural, societal, and familial influences, messages from teachers serve as a critical support for (or threat to) students’ belonging, motivation, and academic identity [11, 37]. Teacher messages of support are particularly important for developing student interest in STEM-related fields. By the time students reach 9$^{th}$ grade, STEM identification, enjoyment, interest, and future utility beliefs are strongly related to perceptions of teacher support [19].

This raises the question: how do teacher messages – in terms of content and form – affect students’ motivation, identity, and sense of belonging? One approach to address this question is to use traditional classroom observation methods [48], including classroom video coding [18]. However, current observation protocols are limited in that they only provide coarse-grained information on teacher practice by pooling observations over long intervals of time, ranging from several minutes to an entire class period. To illustrate, consider the large-scale Measures of Effective Teaching (MET) study [17], where trained annotators scored videos of classroom instruction in six school districts over a two-year period from 2009-2011 using several high-quality video-coding protocols, such as the widely used CLASS observation protocol [32]. In the MET study, trained observers watched 15-minute segments of classroom video, took notes, and provided *one* holistic judgment per CLASS scoring dimension [51]. Thus, a teacher scoring a 2 on the positive emotional climate dimension might be differentiated from one who scored a 5, but the precise behaviors and messages underlying those codes are lost. Whereas this approach might be adequate to obtain a global measure of one or more dimensions of teacher practice, it does not pinpoint fine-grained behaviors, such as specific messages, that teachers can use to reflect upon and potentially alter in their classroom practice. This weakness is particularly significant when subtle and/or infrequent behaviors, such as a callous remark or an unfriendly tone, may be extremely potent for threatening a sense of belonging for some students.

Despite the substantial findings that show the importance of teacher support in creating a normative environment of inclusion and widespread engagement, we lack fine-grained details on the precise teacher behaviors that lead to this outcome. What specific motivational and supportive messages do teachers use in their classrooms to promote (or undermine) students' belonging and academic identity? Providing fine-grained answers to this question is the first step in providing empirically-based, actionable insights for educational researchers, administrators, and teachers to improve pedagogical practices by creating more inclusive and supportive learning environments.

An inherent challenge in developing fine-grained, discourse-based measures is that both the content and form of teacher communication may shape students' perceptions of the classroom environment. Certainly, traditional human coders appraise not just what teachers say (i.e., the verbal content), but how it is said (i.e., paraverbal information). For example, consider the teacher utterance "C'mon miss genius. You are a genius." This statement could be interpreted as either a genuine message of support or as a sarcastic remark. Because paralinguistic cues convey information about the speaker's intentions and mood [12], they can help resolve lexical ambiguities in teacher utterances beyond the words themselves. It could also be the case that students are more sensitive to one of these modalities, raising the question of what matters more: the specific words that teachers say, or the way teachers say them? We address this question by analyzing both the language teachers use along with accompanying paralinguistic cues.

## 1.1 Background and Related Work

*1.1.1 Teacher Support & Student Sense of Belonging.* Teacher support refers to students' perceptions that their teacher cares about them and will help them and is recognized as an essential component for understanding effective instruction [25, 35]. Supportive classroom instruction is related to both student engagement and academic achievement [9, 20, 30]. Along these lines, Osterman proposed two roles that teachers should adopt to support their students: an academic supportive role (teacher as instructor), and a personal supportive role (teacher as person) [28]. Unfortunately, teachers face many obstacles in providing support to their students. High stakes testing policies stemming from initiatives such as No Child Left Behind place the emphasis of teaching on knowledge transmission and test preparation rather than supporting students' learning [28]. Further, some teachers feel neither responsible nor prepared to fill the role of primary motivator and engager of students [34]. Thus, there is a need for strategies and tools to assist teachers in supporting their students' emotional and motivational needs.

Relatedly, student belonging pertains to the extent students feel personally accepted, respected, included, and supported by others in the school social environment [1], and is connected to a range of outcomes, including academic achievement and school engagement [42]. St-Amand et al. [42] found that positive emotions and positive social relations are important for a healthy sense of belonging, both of which are largely determined by social environment. Attaining positive emotions requires feelings of attachment, intimacy, usefulness, and support. Achieving positive social relationships in school requires encouragement, acceptance, support, respect, and warmth [42]. Teacher messaging can directly influence many of these dimensions. Indeed, among the ten themes that influence school belonging found in [1], teacher support, along with positive personal characteristics, was found to be the strongest predictor of students' sense of school belonging. Students feel more willing to engage when they feel cared for, both academically and personally, by their teachers [28]. However, the most effective ways of communicating teacher messages of support are not yet fully understood, especially the role of verbal and paraverbal communication.

*1.1.2 Verbal vs. Paraverbal Communication.* Verbal communication consists of the specific words used during speech, whereas nonverbal communication is defined as behavior of the face, body, or voice minus the linguistic content, in other words, everything but the words [12]. Of particular relevance to this study is the aspect of nonverbal communication called paralanguage (paraverbal vocal cues), which is vocal behavior that occurs with or as a substitute for words [12]. A speaker's paralanguage and words often parallel each other in meaning, with the former modality often resolving linguistic ambiguities or adding information independent of words altogether (e.g., laughing without speaking). Paraverbal cues provide information to listeners about the speakers' motivations and emotions. To illustrate, by examining event-related potentials (ERPs), Zougkou et al. found that listeners appear to rapidly distinguish between controlling (i.e., speech that imposes expectations on how the listener should act) and neutral speech, leading to greater attunement to the former [52]. This finding shows that some paralinguistic cues activate mechanisms in the brain that give preferential treatment to incoming sources of information. Additionally, such cues also convey information about the speakers' affect. For example, in some studies higher pitch, greater pitch range, more loudness, and faster speech rate are associated with joy and elation, and lower pitch, reduced loudness, slower rate, and longer pauses have been linked to sadness [12]. In general, the research on verbal and nonverbal communication in the classroom for over 50 years [10, 40, 41] has indicated the importance of nonverbal communication. Whereas these studies have used traditional analytic techniques, recent work in the field of teacher analytics (reviewed next) has utilized modern methods (e.g., machine learning) to investigate teacher speech modalities in classrooms.

*1.1.3 Teacher Analytics.* Teacher analytics aim to use analytical methods and tools to uncover and understand aspects of effective pedagogy, with a major goal being the provision of insights to teachers to reflect upon and improve their teaching practice. Whereas much of the previous work has focused on automated classification of basic classroom time use (e.g., lecture vs. group work) and features of transactional (dialogic vs. monologic) discourse [5, 8, 14, 15, 26, 44, 50], here we focus on teacher analytics in the support domain and on studies that contrast linguistic and paralinguistic cues in teacher speech.

To this point, Seidel et al. performed a qualitative investigation of instructor language over a semester-long introductory college biology course by examining aspects of *Instructor Talk,* that is, teacher talk that is not content-related [38]. They analyzed more than 600 teacher quotes and identified five emergent categories of Instructor Talk: building the instructor/student relationship, establishing classroom culture, explaining pedagogical choices, sharing personal experiences, and unmasking science. The study, however, did not aim to automate the coding of these talk categories.

A more computationally driven study was conducted by Schlotterbeck et al. who collected audio of teacher classroom sessions using low-cost microphones and used acoustic features (e.g., pitch, energy) to train Random Forest classifiers to classify teacher talk into three categories; *Presenting, Administration,* and *Guiding*, based on the Classroom Observation Protocol for Undergraduate STEM [36]. While the resulting models performed poorly for detecting Administration due to its low prevalence, they were able to predict the presence or absence of Presenting and Guiding with 86% and 83% accuracy respectively. Similarly, Ramakrishnan et al. developed ACORN, a multi-modal machine learning-based system, to analyze videos of school classrooms for classroom climate based on the CLASS observation protocol [32, 33]. Using the MET database's [17] 5,574 CLASS-coded video segments of elementary and middle school classrooms, Random Forest models trained on 200 acoustic features could predict positive and negative climate with an accuracy (Pearson correlation) of 0.36 and 0.41 respectively.

One limitation of the above studies is that they did not consider the content of the teacher utterances, so the question of the relative contribution of verbal vs. paraverbal information remains unanswered. However, outside of the teacher support domain, Donnelly et al. [8], investigated automatic detection of teacher questions from audio recordings using three sets of features: linguistic, acoustic, and contextual. They found that while the models trained on linguistic features outperformed those trained on acoustic and contextual features, combining the feature sets yielded a 5% improvement in accuracy compared to linguistic features alone. A pertinent question is whether a similar additive affect will be observed for teacher talk in the support domain.

## 1.2 Current Study, Novelty, and Research Questions

We leverage a new fine-grained coding scheme for teacher messages related to support for 6$^{th}$ to 8$^{th}$ grade math classes from the MET data (under review). We focus on this age-group because adolescence is a challenging time during which students are beginning to navigate the transition from childhood to adulthood, and for whom, consequently, messages of support are particularly important [1]. Thus, classroom interactions for this demographic of students make for a pertinent testbed to develop systems to measure and improve support-related teacher discourse.

We analyzed linguistic and paralinguistic cues in teacher speech using 2,818 utterances across 156 MET video segments. The teacher utterances were manually transcribed and annotated using the aforementioned coding scheme which builds from existing theory-based protocols assessing teacher behaviors and messages that may affect motivation, engagement, and belonging in classrooms. We trained random forest classifiers to automate the coding using three sets of features: verbal features (language), paraverbal features (acoustics), and a combination of both. We compare the performance of these models along with their top feature importance to gain insight into the relationships between the verbal and paraverbal modalities of teacher speech along with their respective impacts on student perceptions.

We addressed the following research questions: **RQ1.** *What are the relative contributions of the verbal and paraverbal modalities of teacher speech with respect to teacher messages within the support domain;* **RQ2**. *How do verbal and paraverbal models of teacher support correlate with student perceptions of the classroom environment;* **RQ3.** *What specific patterns of language are most relevant for predicting each dimension of support-related teacher messages?*

Our work contributes to the nascent field of teacher analytics, which aims to use analytical methods and tools to provide teachers with insights to reflect upon and improve their pedagogy [39]. As we reviewed above, much of the prior work in this area has focused on supporting content-related instruction, whereas our focus here is specifically on emotional and motivational supports. To our knowledge, ours is the first fine-grained analysis of teacher speech specifically concerned with teacher messages of support.

# 2 METHOD

## 2.1 MET Data Source

The MET study scored videos of instruction in six school districts over a two-year period from 2009-2011 using a high-quality video-coding protocol. We focused our analysis on the year-2 data, where the student perceptions survey included an expanded set of measures. The sample was derived from 156 recorded classroom videos from 73 teachers.

To ensure sufficient variability in classroom environments, our sampling was stratified by the dispersion (SD) in overall Tripod scores (measure of student perceptions of the classroom learning environment, see below), oversampling high and low dispersion class sections. The Tripod survey captures student perceptions on essential elements of instructional practice by asking for their level of agreement with a series of statements related to seven constructs: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate [16]. We focus on the overall composite Tripod score for each teacher (SPS2011ADJ_MATH_COMP1).

Our sampling was clustered by section/teacher, such that only teachers with three or more videos were included, and these three videos were randomly selected. The analyzed sample included data from approximately 1,400 students, of which 52% were male, 47% eligible for free- or reduced-price lunch, 33% Latinx, and 29% Black. Among the 73 teachers, 30% were male, 39% Black, 10% Latinx, averaging 11.3 years of experience on average.

## 2.2 Teacher Speech Coding

The coding scheme builds from existing theory-based protocols assessing teacher behaviors and messages that may affect motivation, engagement, and belonging in classrooms. In particular, it draws from the "connections and applications" dimension of the Mathematics Scan [29], which has been used in prior research to document the relation between teachers' use of relevant real-life examples and students' math attainment value [24]. It also incorporates emerging work documenting verbal and nonverbal instructor behavior indicating a growth mindset. The coding scheme also identifies messages and behaviors postulated to send signals regarding student belonging, value, and identity, drawing from a literature on culturally responsive math instruction [47] and social justice mathematics [21]. More detail about the coding scheme can be found in (under review).

The coding scheme consists of eight codes (Table 1) forming four complementary pairs (Public Praise – Public Admonishment, Autonomy Support – Controlling Language, Learning Mindset Supportive – Learning Mindset Undermining, Strategy Suggestion – Lack of Strategy). They can also be grouped into Supportive and Unsupportive messages (Public Admonishment, Controlling Language, Lack of Strategy, and Learning Mindset Undermining). We further divided the Supportive messages into two subgroups to reflect two aspects of teacher supportive messages: Emotional Support (Public Praise and Learning Mindset Supportive); and Learning Support (Autonomy Support and Strategy Suggestion).

Our coding approach was motivated by given constraints in coding resources and to ensure sufficient variability in the distributions to test our central research questions. We used a nomination coding scheme where coders viewed the 15-minute video clip to identify teacher utterances exhibiting a given discourse property, as opposed to an exhaustive coding process where all speech is transcribed, segmented, and coded. After nominating an utterance, the transcript, timestamp, and code were recorded. A total of 2,818 utterances were independently coded by a trained coder and then checked by an expert coder; see (under review) for details.

## 2.3 Machine Learning

*2.3.1 Random Forest Classifier.* We used a supervised classification approach, specifically a Random Forest Classifier (RF), to model the relationship between teacher speech and the codes described above. We trained the models using teacher-level 8-fold cross validation, which means that all the utterances for a given teacher were either included in the training set or the testing set, but never in both. This approach promotes generalizability to new teachers because it ensures a model is never trained and evaluated on utterances from the same teacher. All RF models were trained using the scikit-learn [31] library implementation with 100 estimators. RF models predict the presence or absence of a given discourse variable by outputting a prediction ranging from 0 to 1. Predictions generated by the models were aggregated up to the video level for analysis.

*2.3.2 Teacher Speech Modality Features.* We investigated three sets of features to train our RF models.

*Verbal Features (N-grams).* The first set of features captures the specific language used by teachers as they communicate with their classrooms. It was derived using a bag of words n-gram approach, which computes counts of words and phrases from the transcribed teacher utterances. We used single words, bigrams (two words), and trigrams (three words) for our n-gram features. We also filtered bigrams and trigrams using a pointwise mutual information (PMI) [4] of 1.0, which indicates the extent to which the words in an n-gram meaningfully co-occur (e.g., "you can") vs. occurring merely by chance. Additionally, we filtered the data to include n-grams that occur with a minimum frequency in the corpus of 1%.

*Paraverbal Features (MFCCs + Speech Production Features).* The second set of features reflects the paralinguistic style in which teachers speak to their classrooms, and it is composed of two subsets. We used the method described in [33] to extract the first 200 Mel-Frequency Cepstral Coefficients (MFCCs) using [46]. MFCCs are a commonly used representation of speech (i.e., they focus on the shape of the spectral band rather than the details) that corresponds to the human auditory system. They are a common workhorse in speech applications, such as speech recognition [45]. MFCCs were extracted across each one second interval of audio for each teacher video and were then averaged across the start and end time for each utterance. We used only the first 12 MFCCs for parsimony, as performance was no better with the full set of 200. The speech production features used were length of utterance in seconds, number of words in utterance, and speech rate (number of words in utterance / length of utterance in seconds).

*Combined.* We combined the verbal and paraverbal feature sets using feature-level fusion.

*2.3.3 Accuracy Metrics.* We used the Area Under the Precision Recall Curve (AUPRC) as our utterance-level accuracy metric because the class base rates were imbalanced (Table 1) and AUPRCs are more robust to class imbalance [13]. Further, an AUPRC corresponding to the base rate provides a chance (guessing) baseline. As another metric of model accuracy, we calculated the video-level Spearman correlations between model predictions and ground truth after aggregating across the coded utterances in each video. We used the Spearman correlation because it is both non-parametric and robust to outliers.

# 3 RESULTS

## 3.1 Incidence of Codes

We found that teachers used Admonishment somewhat more often than Praise (about 44% of evaluative utterances were Praise as opposed to Admonishment). Teachers were also considerably more likely to use Mindset Supportive than Mindset Undermining language (about 82% of mindset-related discourse was Mindset Supportive). Teachers nearly always offered an explicit Strategy Suggestion (91% of strategy-related discourse) as opposed to an obvious/complete lack of strategy. Autonomy Supporting language (46%) and Controlling language occurred in about equal measure.

## 3.2 Model Results Across Modalities

Results of model training are in Table 2. We did not model Lack of Strategy as an independent code due to its extremely low prevalence (1%); however, we still included it for the Unsupportive grouped code. Both modalities did capture meaningful information with average improvements over chance base-rates of 278% and 103% for verbal and paraverbal models, respectively. Accuracies approximately tracked the base rates with the most accurate for Public Admonishment, Praise, and Strategy Suggestion (base rates between 13.5% and 37%) compared to the remaining four codes with base rates less than 8.3%. Both modalities had difficulties in predicting Learning Mindset Undermining language, ostensibly due to its low base rate of 2.8%.

With respect to comparing modalities, the RF models trained on the verbal features achieved higher AUPRCs (mean = .499) and video-level Spearman correlations (mean $\rho$ = .546) than those trained on the paraverbal features for all individual codes (mean AUPRC = .299; mean $\rho$ = .276). Model performances for the grouped codes mimicked those of individual codes with the verbal models

**Table 1: Teacher utterance code descriptions**

| Category | Code | Description | Examples |
|---|---|---|---|
| Emotional Support | Public Praise | Teacher explicitly praises student(s) for ideal or desirable behavior. | “Good, perfect.”<br>“Very good. Because that shows me, [Name] knows how to do this.”<br>“And I like the way this group is paying attention, and they’re following along in the textbook.” |
| | Learning Mindset Supportive | Teacher language that supports growth mindset, purpose and relevance, and social belonging. | “So, if you didn’t get C, it’s alright.”<br>“C’mon miss genius. You are a genius.”<br>“Okay, Imma check it to be sure you did it correctly, and that you are successful.” |
| Learning Support | Autonomy Support | Teacher provides student with a clear choice regarding classroom activity or learning strategy. | “How many of you want one more minute?”<br>“Who would like to explain?”<br>“We’ll do it together, but you can do it in your notebook too.” |
| | Strategy Suggestion | Teacher shares techniques, tools, or tips for learning and understanding material. | “Okay, try to visualize it.”<br>“If it terminates, what does it do? It ends.”<br>“What was the steps? Tell me the steps? Show it to me.” |
| Unsupportive | Public Admonishment | Teacher expresses disapproval of disruptive, inappropriate, or undesirable behavior. | “Pay attention!”<br>“Shh!”<br>“[Name], that was not cute, and it definitely was not funny.” |
| | Lack of Strategy | Teacher withholds or fails to offer a Strategy Suggestion after being implicitly or explicitly prompted. | “I don’t know, you need to find out.”<br>“Check your notes.”<br>“So that’s something that you kinda gotta just remember.” |
| | Controlling Language | Teacher emphasizes a lack of student autonomy and agency, often indicating there is only one right way to complete a task. | “Hurry up, hurry up, chop chop!”<br>“No! Get the brown – you need to use what I’m telling you, [Name].”<br>“Just don’t write so big!” |
| | Learning Mindset Undermining | Teacher language that undermines growth mindset, purpose and relevance, and social belonging. | “Ok, you’re a little behind everybody.”<br>“Who was your fifth-grade teacher last year?”<br>“So you need to pass the test is what you gotta do.” |

outperforming the paraverbal models. However, grouped code models outperformed those of the individual codes across the board, likely due to the better balance of class distributions. Meng’s test of dependent overlapping correlations indicated that the differences were highly significant for all the variables ($p < .01$ for all) with a somewhat smaller difference ($\rho$ of .605 (verbal) and .527 (paraverbal), $p < .05$) for Strategy Suggestion. While the paraverbal models themselves were not very performant in direct comparison to more verbal models, a more pertinent question is whether it can offer any reliable signal whatsoever in the support domain. The results indicate that this was achieved, and thus, it would be premature to dismiss paraverbal cues in this area of research. We also experimented with combining the verbal and paraverbal features but the results (mean $\rho = .542$) did not outperform the verbal modality (mean $\rho = .546$) alone in these data (details not shown).

## 3.3 Tripod Correlations

We analyzed the video-level Spearman correlations between model predictions and students’ perception of the classroom environment (from Tripod – Table 3).

Of the eight original codes, only three were significantly correlated with the Tripod outcomes: Public Admonishment, Public Praise, and Learning Mindset Supportive. Interestingly, while RF Verbal model predictions showed similar correlations to ground truth with Public Admonishment and Public Praise, it was the RF Paraverbal model that was closest to ground truth for Learning Mindset Supportive. This supports the hypothesis that paraverbal information is indeed an important factor in student perceptions of their teachers. As for the grouped codes, only the Unsupportive and Emotional Support ground truth codes significantly correlated with Tripod scores, a pattern replicated for the verbal model and partly (i.e., for Emotional Support) for the paraverbal model.

## 3.4 Analysis of Language Features

To gain insight on the specific language associated with support-related teacher messages, we investigated the top 10 most important n-grams for each code of the RF Verbal models. We identified the top n-grams by averaging n-gram importances across each fold of 8-fold cross-validation. The results (Table 4) reveal patterns that demonstrate both the complementary nature of the codes as well

**Table 2: Random Forest Model AUPRCs and Video-level Spearman Correlations of Model Predictions and Ground Truth**

| Code | Base Rate (N=2,818) | AUPRC Verbal | Paraverbal | Spearman Corr. Verbal | Paraverbal |
|---|---|---|---|---|---|
| **Individual Codes** | | | | | |
| Public Admonishment | 37.0% | .910 | .611 | .878 *** | .453 *** |
| Public Praise | 24.7% | .967 | .586 | .939 *** | .558 *** |
| Strategy Suggestion | 13.4% | .612 | .477 | .605 *** | .527 *** |
| Controlling Language | 8.3% | .251 | .133 | .466 *** | .215 *** |
| Autonomy Support | 6.8% | .477 | .128 | .459 *** | .164 * |
| Learning Mindset Supportive | 6.0% | .230 | .125 | .376 *** | .221 ** |
| Learning Mindset Undermining | 2.8% | .047 | .030 | .099 | -.206 |
| Lack of Strategy Suggestion | 1.0% | - | - | - | - |
| **Mean** | | .499 | .299 | .546 | .276 |
| **Grouped Codes** | | | | | |
| Unsupportive | 49.1% | .943 | .677 | .882 *** | .458 *** |
| Emotional Support | 30.7% | .911 | .556 | .881 *** | .429 *** |
| Learning Support | 20.2% | .738 | .532 | .717 *** | .586 *** |

$p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***

**Table 3: Video-level Spearman correlations between model predictions and overall Tripod scores (N=156 Videos)**

| Code | Ground Truth | RF Verbal | RF Paraverbal |
|---|---|---|---|
| **Individual Codes** | | | |
| Public Admonishment | -.197 * | -.208 ** | -.129 |
| Public Praise | .226 ** | .204 * | .022 |
| Strategy Suggestion | .010 | .051 | .128 |
| Controlling Language | -.150 | -.062 | .106 |
| Autonomy Support | .145 | .009 | .082 |
| Learning Mindset Supportive | .161 * | .080 | .177 * |
| Learning Mindset Undermining | -.064 | -.015 | .097 |
| **Grouped Codes** | | | |
| Unsupportive | -.276 *** | -.196 * | -.212 ** |
| Emotional Support | .288 *** | .222 ** | .096 |
| Learning Support | .095 | .054 | .105 |

$p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***

as the constructs they are designed to capture. For example, Public Admonishment and Public Praise shared four top n-grams: *shh, good, [name],* and *you*. While *shh* occurred 330 times (31.6%) in Public Admonishment utterances, it occurred exactly once for Public Praise. Thus, it's presence strongly predicts the former and its absence the latter. Similarly, *good* occurred 509 time (73.2%) for Public Praise utterances and only six times (0.6%) for Public Admonishment. Less obvious is that the occurrence of *[name]* (i.e., the teacher said the name of a student) was positively correlated (Spearman) with Public Admonishment (0.265) and negatively with Public Praise (-0.102), suggesting that when teachers say student names it is likely because that student is misbehaving.

The terms *you-want* ("How many of *you want* one more minute?") and *you-can* ("We'll do it together, but *you can* do it in your notebook, too."), which present students with a choice, were strong predictors of Autonomy Support, whereas commanding statements like "Let's *go* quickly!" and "Just don't *write* so big!" indicated Controlling Language. For Learning Mindset Support, we found terms that indicate relationship, like *you* and *I* ("*I* want *you* guys to be 100%."), *help* ("You're going to *help* each other"), and *give* ("[Name], let's *give* everyone a chance."). Learning Mindset Undermining, however, was predicted by terms that position the student as an isolated learner, such as *you, know, think,* and *do* ("*You* need to *know* that tomorrow and *you* can't ask me anything

**Table 4: Top 10 Correlated N-grams**

| Code | Top 10 N-grams |
|---|---|
| Public Admonishment | *shh, good, [name], please, it, the, you, stop, that, very* |
| Public Praise | *good, very, very-good, perfect, excellent, shh, you, job, to, [name]* |
| Strategy Suggestion | *the, what, so, is, to, a, if, remember, you, it* |
| Controlling Language | *up, go, write, you, do, no, it, the, on, [name]* |
| Autonomy Support | *you-want, you-can, can, you, want, or, it, to, the, would* |
| Learning Mindset Supportive | *help, you, it, a, that, I, they, if, to, give* |
| Learning Mindset Undermining | *that, you, know, who, the, only, to, think, for, do* |

during the test", "Ok, *you're* a little behind everybody", "*Think*!"). Overall, these patterns confirm that the RF models are successfully capturing the teacher language that the coding scheme is meant to target.

## 4 DISCUSSION

Teacher support for students' socio-emotional needs in the classroom is essential for effective teaching and learning, as teacher messages influence student perceptions of school belonging, motivation, and academic identity. Thus, developing automated methods to detect fine-grained discourse features has important implications for both educational researchers and teacher professional development. Our current work serves as a first step towards developing an automated system to measure constructs of teacher support-related discourse from recorded classroom audio.

### 4.1 Main Findings

Our first research question (RQ1) was to determine the relative contributions of the verbal and paraverbal modalities of teacher speech with respect to teacher messages in the support domain. We found that verbal cues outperformed paraverbal cues for all discourse features and their combinations. This is unsurprising for codes such as *Public Admonishment* and *Praise*, as both codes are evidenced by the presence or absence of particular words (e.g., *"shh", "good"*). However, the paraverbal models out-performed chance guessing for all codes except *Learning Mindset Undermining*, which was also poorly predicted based on verbal cues. This shows that the paraverbal modality of teacher speech conveys a detectable signal indicative of teacher support-related messaging. Despite the predictive power of both modalities of teacher speech, combining them yielded no improvement in model performance. It should be emphasized that our analysis was performed using manually segmented and transcribed teacher utterances. Thus, while the paraverbal modality added no value to the high-quality verbal information in this study, results might change for analysis with automatic speech recognition (ASR)-based transcripts, which are prone to erroneous transcriptions.

We then investigated how the verbal and paraverbal cues correlated with student perceptions of classroom environment via their Tripod scores (RQ2). We reasoned that obtaining an external measure independent of the training data was necessary to validate the overall approach. Three of the eight (human-coded) individual discourse codes (*Public Admonishment*, *Public Praise*, and *Learning Mindset Supportive*) correlated with the overall Tripod score, a pattern replicated by two of the verbal models (*Public Admonishment* and *Praise*) and one paraverbal model (*Learning Mindset Supportive*). Although determining whether students themselves agree with our coding scheme is an important area of future work, the correlation between these three codes and the overall Tripod score indicates some degree of alignment We also found that both models' predictions of *Unsupportive* grouped codes correlated with Tripod scores, suggesting that student perceptions are particularly sensitive to unsupportive messaging for both verbal and paraverbal cues. This finding indicates that it is possible to train models to predict student perceptions of classroom environment from MFCCs extracted from audio recordings without the need for long-term storage of classroom audio or transcriptions, a relevant finding for protecting the privacy of students and teachers alike.

Finally, we investigated the specific patterns of language that are most relevant for predicting each dimension of support-related teacher messages (RQ3). The top 10 important n-grams identified by the Random Forest verbal models helped identify relevant examples of teacher utterances for each code (e.g., the word "*give*" for *Learning Supportive Mindset* in the utterance *"[Name], let's give everyone a chance"*). This analysis can be used to ground codes in salient example utterances, which can then be used to present relevant examples for teachers to reflect upon. Additionally, providing teachers with certain words and phrases that commonly occur with a given discourse feature can help them appropriately adjust the language they use in their lessons.

### 4.2 Applications, Limitations, and Future Work

A primary goal of teacher analytics is providing educators with feedback for professional development. It is important to mention that automated feedback is not meant to replace traditional methods of teacher feedback by peers and observers. However, through ease of accessibility and relatively low costs, automated methods can alleviate the cost-prohibitive constraints of traditional approaches. Fine-grained feedback systems can also provide the content and platform for peer feedback and professional learning communities [3]. While there is ongoing work to automate feedback from global protocols that target teacher support (e.g., CLASS), the feedback generated is not specific to precise teacher behaviors. The fine-grained approach used in this study addresses this weakness by targeting individual utterances of teacher speech. Consequently, a key application of the current work is generating and providing teachers with formative feedback as a tool for self-reflection on the specific language used in their support-related discourse.

Feedback generated from this fine-grained approach can be subsequently aggregated up to the class level, providing teachers with both general and specific information about their support-related messaging. These two granularities of feedback provide teachers with high-level notions of their supportive discourse as well as specific examples of their speech, thus allowing them to ground the feedback in their actual spoken words. For example, if a teacher wanted to reduce the amount of Controlling Language they use, they could obtain a sample of utterances with high predicted probabilities for that code and its complement, Autonomy Support. This would give them concrete examples of both speech to avoid (*"Hurry up, hurry up, chop chop"*) and speech to promote (*"How many of you want one more minute?"*), ideally making it easier to incorporate feedback into their teaching practice. However, we recognize that this work is only a first step towards such a system, as our study was subject to limitations.

First, our dataset was comprised of manually transcribed teacher utterances resulting in highly accurate representations of the verbal modality of teacher speech. Hence, our findings may not replicate for inherently noisy ASR-generated datasets from real-world classrooms, which are necessary to provide automated teacher feedback [6]. We did not use ASR in this work due to data access restrictions but replication with ASR transcripts on different data sets would be desirable. A second limitation was that our dataset was composed of a set of nominated utterances, which requires human intervention. A fully automated system would have to work with *all* teacher utterances, even those that do not fall within the support domain. That said, it should be noted that the present goal was not to develop a fully automated system but to conduct basic research into patterns of teacher discourse, an essential step in building such a system. Another limitation was that we restricted our modeling approach to Random Forest classifiers for the sake of interpretability. Future work should investigate other, state-of-the-art modeling approaches. A final limitation was that we were unable to reliably attach student IDs to specific teacher utterances, thereby limiting our ability to investigate teacher discourse directed at particular student groups, especially those with stigmatized identities with respect to gender, race/ethnicity, and socioeconomic status. This is particularly relevant in the case of math as such students face cultural stereotypes that undermine both their innate abilities and their views on the importance of math for their lives [22, 23].

Addressing the above limitations with a different corpus is one important aspect of future work. Another direction is to assess whether middle school math students themselves would agree with the coding of support-related discourse used in the present study. Researchers could use an experience sampling approach during instruction to collect student perceptions of support, or they could retrospectively collect these data after a class. This would be an important step towards including the students themselves as stakeholders when building systems designed to capture some aspect of the teacher-student relationship.

### 4.3 Concluding Remarks

We investigated teacher discourse in the support domain using coded video segments of middle school math classes. We found that while the verbal modality of speech was more predictive, paraverbal cues also contained enough signal to predict a majority of the codes over chance. The present work adds to broader research on teacher analytics by investigating how teacher practices relate to students' psychological outcomes, while also providing actionable information to help teachers improve their own practices, which should have positive downstream consequences for students.

## ACKNOWLEDGMENTS

This research is supported by the Student Experience Research Network (SERN) and the Bill and Melinda Gates Foundation. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

## REFERENCES

[1] Allen, K. *et al.* 2018. What Schools Need to Know About Fostering School Belonging: a Meta-analysis. Educational Psychology Review. 30, 1 (Mar. 2018), 1–34. DOI:https://doi.org/10.1007/s10648-016-9389-8.
[2] Allen, K.A. and Bowles, T. 2012. Belonging as a Guiding Principle in the Education of Adolescents. Australian Journal of Educational & Developmental Psychology. 12, (2012), 108–119.
[3] Caughlan, S. *et al.* 2013. English Teacher Candidates Developing Dialogically Organized Instructional Practices. Research in the Teaching of English. 47, 3 (2013), 212–246.
[4] Church, K.W. and Hanks, P. 1989. Word association norms, mutual information, and lexicography. Proceedings of the 27th annual meeting on Association for Computational Linguistics - (Morristown, NJ, USA, 1989), 76–83.
[5] Cook, C. *et al.* 2018. An open vocabulary approach for estimating teacher use of authentic questions in classroom discourse. Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018. (2018), 116–126.
[6] D'Mello, S.K. *et al.* 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction. (2015), 557–566. DOI:https://doi.org/10.1145/2818346.2830602.
[7] Degelsmith, A.B. 2000. The relationship between students' sense of belonging in school and academic outcomes. University of Georgia.
[8] Donnelly, P. *et al.* 2016. Automatic teacher modeling from live classroom audio. UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. (2016), 45–53. DOI:https://doi.org/10.1145/2930238.2930250.
[9] Furrer, C. and Skinner, E. 2003. Sense of relatedness as a factor in children's academic engagement and performance. Journal of Educational Psychology. 95, 1 (Mar. 2003), 148–162. DOI:https://doi.org/10.1037/0022-0663.95.1.148.
[10] Galloway, C. 1968. Nonverbal communication. Theory Into Practice. 7, 5 (Dec. 1968), 172–175. DOI:https://doi.org/10.1080/00405846809542151.
[11] Gray, D.L.L. *et al.* 2018. Black and Belonging at School: A Case for Interpersonal, Instructional, and Institutional Opportunity Structures. Educational Psychologist. 53, 2 (2018), 97–113. DOI:https://doi.org/10.1080/00461520.2017.1421466.
[12] Hall, J.A. *et al.* 2019. Nonverbal Communication. Annual Review of Psychology. 70, 1 (Jan. 2019), 271–294. DOI:https://doi.org/10.1146/annurev-psych-010418-103145.
[13] Jeni, L.A. *et al.* 2013. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (Sep. 2013), 245–251.
[14] Jensen, E. *et al.* 2021. A deep transfer learning approach to modeling teacher discourse in the classroom. ACM International Conference Proceeding Series. (2021), 302–312. DOI:https://doi.org/10.1145/3448139.3448168.
[15] Jensen, E. *et al.* 2020. Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning. Conference on Human Factors in Computing Systems - Proceedings. (2020), 1–13. DOI:https://doi.org/10.1145/3313831.3376418.
[16] Kane, T. and Staiger, D. 2012. Gathering feedback for teaching. Measures of Effective Teaching Project: Bill and . . .. (2012).
[17] Kane, T.J. *et al.* 2013. Have We Identified Effective Teachers? Validating measure of effective teaching using random assignment. Research Paper: MET Project. (2013), 52.
[18] Kelly, S. *et al.* 2020. Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. Education Policy Analysis Archives. 28, (2020), 1–34. DOI:https://doi.org/10.14507/EPAA.28.5012.
[19] Kelly, S. and Zhang, Y. 2016. Teacher Support and Engagement in Math and Science: Evidence from the High School Longitudinal Study. The High School Journal. 99, 2 (2016), 141–165. DOI:https://doi.org/10.1353/hsj.2016.0005.

[20] Klem, A.M. and Connell, J.P. 2004. Relationships matter: Linking teacher support to student engagement and achievement. Journal of School Health. 74, 7 (2004), 262–273. DOI:https://doi.org/10.1111/j.1746-1561.2004.tb08283.x.

[21] Kokka, K. 2015. Addressing Dilemmas of Social Justice Mathematics through Collaboration of Students, Educators, and Researchers. Educational Considerations. 43, 1 (Oct. 2015). DOI:https://doi.org/10.4148/0146-9282.1031.

[22] Martin, D. 2012. Learning Mathematics While Black. The Journal of Educational Foundations. 26, 1/2 (2012), 47.

[23] Matthews, J.S. *et al.* 2014. Academic Identity Formation and Motivation Among Ethnic Minority Adolescents: The Role of the "Self" Between Internal and External Perceptions of Identity. Child Development. 85, 6 (2014), 2355–2373. DOI:https://doi.org/10.1111/cdev.12318.

[24] Matthews, J.S. 2018. When am I ever going to use this in the real world? Cognitive flexibility and urban adolescents' negotiation of the value of mathematics. Journal of Educational Psychology. 110, 5 (Jul. 2018), 726–746. DOI:https://doi.org/10.1037/edu0000242.

[25] Nilsen, T. and Gustafsson, J.-E. 2016. Teacher Quality, Instructional Quality and Student Outcomes: Relationships Across Countries, Cohorts and Time . Springer International Publishing .

[26] Olney, A.M. *et al.* 2017. Assessing the dialogic properties of classroom discourse: Proportion models for imbalanced classes. Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017. (2017), 162–167.

[27] Osterman, K.F. 2000. Students' Need for Belonging in the School Community. Review of Educational Research. 70, 3 (2000), 323. DOI:https://doi.org/10.2307/1170786.

[28] Osterman, K.F. 2010. Teacher Practice and Students' Sense of Belonging. International Research Handbook on Values Education and Student Wellbeing. T. Lovat *et al.*, eds. Springer Netherlands. 239–260.

[29] Ottmar, E.R. *et al.* 2015. Mathematical Knowledge for Teaching, Standards-Based Mathematics Teaching Practices, and Student Achievement in the Context of the Responsive Classroom Approach. American Educational Research Journal. 52, 4 (Aug. 2015), 787–821. DOI:https://doi.org/10.3102/0002831215579484.

[30] Patrick, H. *et al.* 2007. Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. Journal of Educational Psychology. 99, 1 (2007), 83–98. DOI:https://doi.org/10.1037/0022-0663.99.1.83.

[31] Pedregosa, F. *et al.* 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12, (2011), 2825–2830. DOI:https://doi.org/10.5555/1953048.2078195.

[32] Pianta, R.C. *et al.* 2008. Classroom assessment scoring system manual, K-3. Reference & Research Book News. Ringgold, Inc.

[33] Ramakrishnan, A. *et al.* 2021. Toward Automated Classroom Observation: Multimodal Machine Learning to Estimate CLASS Positive Climate and Negative Climate. IEEE Transactions on Affective Computing. (2021), 1–1. DOI:https://doi.org/10.1109/TAFFC.2021.3059209.

[34] Ryan, R.M. and Brown, K.W. 2005. Legislating competence: High-stakes testing policies and their relations with psychological theories and research. (2005).

[35] Scheerens, J. 2014. School, teaching, and system effectiveness: some comments on three state-of-the-art reviews. School Effectiveness and School Improvement. 25, 2 (Apr. 2014), 282–290. DOI:https://doi.org/10.1080/09243453.2014.885453.

[36] Schlotterbeck, D. *et al.* 2021. What classroom audio tells about teaching: A cost-effective approach for detection of teaching practices using spectral audio features. ACM International Conference Proceeding Series. (2021), 132–140. DOI:https://doi.org/10.1145/3448139.3448152.

[37] Schmidt, J.A. *et al.* 2019. Why are we learning this? Using mixed methods to understand teachers' relevance statements and how they shape middle school students' perceptions of science utility. Contemporary Educational Psychology. 57, August 2018 (2019), 9–31. DOI:https://doi.org/10.1016/j.cedpsych.2018.08.005.

[38] Seidel, S.B. *et al.* 2015. Beyond the biology: A systematic investigation of noncontent instructor talk in an introductory biology course. CBE Life Sciences Education. 14, 4 (2015), 1–14. DOI:https://doi.org/10.1187/cbe.15-03-0049.

[39] Sergis, S. and Sampson, D.G. 2017. Teaching and learning analytics to support teacher inquiry: A systematic literature review. Learning analytics: Fundaments, applications, and trends. (2017), 25–63.

[40] Simpson, A.W. and Erickson, M.T. 1983. Teachers' Verbal and Nonverbal Communication Patterns as a Function of Teacher Race, Student Gender, and Student Race. American Educational Research Journal. 20, 2 (Jan. 1983), 183–198. DOI:https://doi.org/10.3102/00028312020002183.

[41] Smith, H.A. 1979. Nonverbal Communication in Teaching. Review of Educational Research. 49, 4 (Dec. 1979), 631–672. DOI:https://doi.org/10.3102/00346543049004631.

[42] St-Amand, J. *et al.* 2017. Sense of Belonging at School: Defining Attributes, Determinants, and Sustaining Strategies. IAFOR Journal of Education. 5, 2 (Sep. 2017). DOI:https://doi.org/10.22492/ije.5.2.05.

[43] Steele, C.M. 1997. A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance. American Psychologist. 52, 6 (1997), 613–629. DOI:https://doi.org/10.1037/0003-066X.52.6.613.

[44] Suresh, A. *et al.* 2019. Automating Analysis and Feedback to Improve Mathematics Teachers' Classroom Discourse. Proceedings of the AAAI Conference on Artificial Intelligence. 33, (Jul. 2019), 9721–9728. DOI:https://doi.org/10.1609/aaai.v33i01.33019721.

[45] Tiwari, V. 2010. MFCC and its applications in speaker recognition. International Journal on Emerging Technologies. 1, 1 (2010), 19–22.

[46] tuneR: Analysis of Music and Speech: 2018. https://cran.r-project.org/package=tuneR.

[47] Turner, E.E. and Drake, C. 2016. A Review of Research on Prospective Teachers' Learning About Children's Mathematical Thinking and Cultural Funds of Knowledge. Journal of Teacher Education. 67, 1 (Jan. 2016), 32–46. DOI:https://doi.org/10.1177/0022487115597476.

[48] Wallace, T.L. *et al.* 2016. What Can Student Perception Surveys Tell Us About Teaching? Empirically Testing the Underlying Structure of the Tripod Student Perception Survey. American Educational Research Journal. 53, 6 (Dec. 2016), 1834–1868. DOI:https://doi.org/10.3102/0002831216671864.

[49] Walton, G.M. and Cohen, G.L. 2007. A question of belonging: Race, social fit, and achievement. Journal of Personality and Social Psychology. 92, 1 (2007), 82–96. DOI:https://doi.org/10.1037/0022-3514.92.1.82.

[50] Wang, Z. *et al.* 2014. Automatic classification of activities in classroom discourse. Computers and Education. 78, (2014), 115–123. DOI:https://doi.org/10.1016/j.compedu.2014.05.010.

[51] White, M. and Rowan, B. 2014. User guide to Measures of Effective Teaching longitudinal database (MET LDB).

[52] Zougkou, K. *et al.* 2017. ERP correlates of motivating voices: Quality of motivation and time-course matters. Social Cognitive and Affective Neuroscience. 12, 10 (2017), 1687–1700. DOI:https://doi.org/10.1093/scan/nsx064.