Sit Down *Now*: How Teachers' Language Reveals the Dynamics of Classroom Management Practices*

Mei Tan† Dorottya Demszky‡

September 6, 2023

Abstract

Teachers' attitudes and classroom management practices critically affect students' academic and behavioral outcomes, contributing to the persistent issue of racial disparities in school discipline. Yet, identifying and improving classroom management at scale is challenging, as existing methods require expensive classroom observations by experts. We apply natural language processing methods to elementary math classroom transcripts to computationally measure the frequency of teachers' classroom management language in instructional dialogue and the degree to which such language is reflective of punitive attitudes. We find that the frequency and punitiveness of classroom management language show strong and systematic correlations with human-rated observational measures of instructional quality, student and teacher perceptions of classroom climate, and student academic outcomes. Our analyses reveal racial disparities and patterns of escalation in classroom management language. We find that classrooms with higher proportions of Black students experience more frequent and more punitive classroom management. The frequency and punitiveness of classroom management language escalate over time during observations, and these escalations occur more severely for classrooms with higher proportions of Black students. Our results demonstrate the potential of automated measures and position everyday classroom management interactions as a critical site of intervention for addressing racial disparities, preventing escalation, and reducing punitive attitudes.

Keywords: classroom management, artificial intelligence, natural language processing, instructional practices, observational research, research methodology, equity

*The authors are grateful to Sanne Smith, the Stanford Education Science '23 Masters Cohort, Heather Hill, Jing Liu, Alvin Pearman, and Juergen Dieber for their invaluable feedback. We also appreciate the hard work of our teacher annotators.

†Mei Tan (corresponding author, mxtan@stanford.edu) is a PhD student at the Stanford Graduate School of Education.

‡Dorottya (Dora) Demszky (ddemszky@stanford.edu) is an Assistant Professor at the Stanford Graduate School of Education.

1 Introduction

Reducing both the overall use of exclusionary discipline and racial disciplinary disparities remains a persistent challenge in U.S. schools (Barrett et al., 2021; J. Liu, Hayes, & Gershenson, 2022; Skiba et al., 2011). Addressing these problems requires better understanding and improving classroom processes. Teachers' behaviors and punitive attitudes contribute to inequitable responses to student behaviors and negative disciplinary outcomes (Chin et al., 2020; Gregory et al., 2010). Classroom management is a key area of intervention, affecting school climate and behavioral outcomes for students (Simonsen et al., 2008; Ingersoll & Smith, 2003; Stronge et al., 2011; Shinn et al., 1987). Effective classroom management significantly decreases disruptive behavior, increases student engagement, and increases academic achievement (Korpershoek et al., 2016; Gage et al., 2018; Lekwa et al., 2019). Across several commonly used classroom observation tools, the classroom management dimension is most consistently and strongly related to teachers' value-added scores across instruments, subjects, and grade levels (Gill et al., 2016). Yet, teachers commonly report feeling stressed and under-prepared to manage classrooms, expressing ongoing concerns about student behavior and frustrations with insufficient support (Reinke, Stormont, et al., 2011; Reinke et al., 2013).

The current predominant approach to evaluating and improving classroom management practice relies on teacher observation. Trained observers attend or watch recordings of classrooms and leverage observational instruments to assess competencies and identify opportunities for professional development (J. Cohen & Goldhaber, 2016). Observational tools such as the Classroom Assessment and Scoring System (CLASS) help score the quality of teacher-student interactions in the classroom (Pianta et al., 2008), providing insights for teachers and administrators. While observational approaches generate detailed assessments of teacher performance, these methods are resource intensive (Wallace et al., 2020; Archer et al., 2016), limiting the frequency and regularity of implementation.

Recent research has demonstrated the potential for developing automated techniques

to complement manual classroom observational processes. Advances in natural language processing has spurred the development of computational approaches to analyzing instructional dialogue. These approaches aim to discover patterns of effective instruction (J. Liu & Cohen, 2021) and identify teachers' use of instructional discourse practices, such as authentic questions (Kelly et al., 2018), uptake of student ideas (Samei et al., 2014; Stone et al., 2019; Demszky et al., 2021), and growth mindset supportive language (Hunkins et al., 2022). While no measures have yet been developed to analyze classroom management practices in instructional dialog, recent studies have applied linguistic analyses to text data in office disciplinary referral records. Markowitz et al. (2023) found that teachers wrote longer descriptions and included more negative emotion when disciplining Black students compared to White students.

We extend this line of work by examining the language of classroom management in teacher-student interactions. Using a large dataset of transcripts from elementary math classroom observations (Demszky & Hill, 2022), our study leverages NLP methods to computationally identify dimensions of classroom management in teacher language. We develop automated measures that identify classroom management language and assess the degree to which such language is reflective of punitive attitudes. We apply these measures to investigate the following research questions:

- RQ1: How does the frequency and punitiveness of classroom management language correlate with observation scores of instruction quality, teacher and student perceptions of classroom climate, and student learning outcomes?

- RQ2: How do teacher, student, and classroom characteristics correlate with the frequency and punitiveness of classroom management language?

- RQ3: How does the frequency and punitiveness of classroom management language change over the course of a classroom observation?

By examining language patterns at scale, our study contributes the first quantitative

analysis of how and when teachers manage their classrooms. The richness of our data allows us to further provide insight into contributing factors and the effects of more frequent and punitive classroom management practice. We hope these insights will help identify key areas of intervention to prevent disciplinary escalation, reduce the use of punitive practices, and eradicate racial disciplinary disparities through improved professional learning for teachers. We release our annotation schema, dataset, and automated measures for other researchers to use and build on.

2 Background on Classroom Management

Defined as actions taken by a teacher to create and maintain an environment conducive to successful instruction (Evertson et al., 2006; Brophy, 2006), classroom management is a powerful component of classroom climate, behavior, engagement, and the quality of student learning. Classroom management is complex and multifaceted—it involves managing classroom time, space, student behavior, and instructional strategies (McLeod et al., 2003). Teachers manage classrooms by establishing rules and procedures, maintaining student attention and engagement, behavior modification, counseling, and administration of supplies or group work. Valid measures of classroom management are vital to specifying the dimensions of effective practice, assessing current teacher behaviors, and supporting professional development.

Various observational tools have been developed to measure classroom management skills, including practitioner-friendly websites and checklists, as well as assessment instruments with established psychometric properties. One frequently cited tool is the Classroom Assessment and Scoring System (CLASS), an instrument developed to analyze the quality of teacher-student interactions in the classroom (Pianta et al., 2008). Across the categories of "instructional support", "emotional support", and "classroom organization", it assesses 11 dimensions of practice through checklists and teacher observations at defined

intervals. These dimensions include "behavior management", "productivity", "positive climate", "teacher sensitivity", and "regard for student perspectives". Other observation tools embed dimensions of classroom management in evaluations developed for specific content areas or pedagogical frameworks. For example, the Mathematical Quality of Instruction (MQI) instrument captures five subject-specific dimensions of instruction, one of which captures whether instructional time is spent on activities that do not develop mathematical ideas, such as transitions or discipline (Hill et al., 2008). The Culturally Responsive Instruction Observation Protocol (CRIOP) sets standards for implementing culturally responsive instruction practices and assesses classroom climate and teachers' ethic of care in addition to the quality of curricula, assessments, and instructional techniques (Powell et al., 2016).

One critical dimension of classroom management reflected in these measures is the distribution of time dedicated to academic instruction and engagement. Effective classroom management is characterized by maximizing the time that students are engaged with academic content (Anderson et al., 1980) and minimizing time spent getting organized (Brophy, 2006). A measure of students' academically engaged time is derived in part by subtracting the amount of time spent on classroom management tasks from the total instructional time (Gettinger & Walter, 2012).

Observational instruments additionally address a complex affective dimension of classroom management related to the measurement of classroom climate resulting from teacher attitudes underlying classroom management actions. Studies on classroom management have linked student misbehavior to teachers' undesired attitudes and threatening environments (Cummings, 2000; Mitchell & Bradshaw, 2013). These attitudes reflect historical behavioral approaches to classroom management and systems with an over-reliance on punitive methods of control (Landrum & Kauffman, 2013). Punitive practices favor consequences to regain control in response to disruption and may involve threatening, shaming, or displaying negative affect toward students. Such practices have been proven ineffective (Van Acker et al., 1996), and effective classroom management is characterized by positive strategies for responding

to inappropriate behavior (Reinke, Herman, & Sprick, 2011). The shift away from punitive practices is characterized by overlapping frameworks and definitions. Restorative practices represent a shift away from control mindsets and toward relational practices and collaborative mindsets (Smith et al., 2015; Buckmaster, 2016). Culturally responsive management involves building caring classrooms sensitive to students' backgrounds and their broader social contexts (Weinstein et al., 2004). Interventions replacing punitive mindsets with empathic mindsets encourage teachers to understand and value students' perspectives and to help students to appropriately conduct themselves in the classroom (Okonofua, Paunesku, & Walton, 2016). Similarly, approaches from positive psychology and autonomy-supportive practices replace compliance-oriented management strategies with those scaffolding students' self-regulatory capacities and engaging students in open communication (Bear et al., 2017; Wallace et al., 2014).

We draw on this body of literature in education and social psychology to develop computational measures that quantify the frequency of and punitive attitudes reflected in classroom management language.

3 Data

The data for this study comprises 1,625 transcripts of math classroom observations collected by the National Center for Teacher Effectiveness (NCTE) between 2010 and 2013 (Kane et al., 2015; Demszky & Hill, 2022). The transcripts capture data from 45-60 minute-long observations from 4th and 5th-grade elementary classrooms. This sample represents 317 teachers across four school districts and 53 schools in the US that serve largely low-income students of color. The dataset includes classroom observation ratings using the Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008) and the Mathematical Quality of Instruction (MQI) (Hill et al., 2008) instruments, teacher background data, value-added scores, student administrative and demographic information, and teacher and student

questionnaire responses on perceptions of classroom climate. Though we acknowledge that the data may not fully represent teachers' typical handling of student behavior due to the presence of cameras and observers during data collection, the transcripts capture identifiable characteristics of classroom management language. Our analyses draw on the 286,561 teacher utterances captured by the transcript data. The teacher utterances in the NCTE dataset are units of speech pre-segmented in the transcription process and contain 29 words on average. We report detailed features of the dataset in Appendix A.

4 Methods

We apply a traditional machine learning paradigm, leveraging human expertise to create a labeled dataset that is then used to develop and validate automated measures. First, we define an annotation scheme for classroom management language in teacher utterances (Section 4.1). We train annotators to apply the scheme and label a small subset of the dataset (Section 4.2). Next, we train machine learning models using the annotated data, assessing accuracy and reliability (Section 4.3). We then apply the models to the full dataset to predict automated measures of classroom management language (Section 4.4). In subsequent sections, we perform downstream analyses of the relationships between classroom management language and instructional context.

4.1 Schema Definition

We develop an annotation scheme by conducting a qualitative review of the classroom management literature to identify practices observable in discourse. First, we define the annotation task of **identifying classroom management language**. We draw on the distinction between academic and procedural language (J. Liu & Cohen, 2021) and define our criteria for classroom management language as language unrelated to academic content that involves the management of the instructional environment. Next, we define the annotation task of

classifying the attitudes underlying classroom management language as **punitive, positive, and neutral**. The specification for this complex dimension involves borrowing terminology from the literature on discipline, psychology, and classroom management. Our definition for positive attitudes synthesizes and consolidates literature on postive psychology (Bear et al., 2017), restorative practices (Smith et al., 2015), empathic mindsets (Okonofua, Paunesku, & Walton, 2016), culturally-responsive management (Weinstein et al., 2004), and autonomy-supportive management (Reeve, 2016). An utterance is punitive if it contains language that seeks to shame, threaten, punish, or otherwise take control of the classroom. An utterance is positive if it contains non-pressuring, informational language that values and seeks to understand student perspectives, supports student growth and agency, and maintains positive relationships with students. An utterance is neutral if it displays no indication of either attitude. Table 1 details definitions and examples for each dimension of the annotation scheme. We additionally define annotation schema for identifying specific talk moves, such as praise and explanatory rationale—discussing the talk moves is beyond the scope of this work, but we provide brief definitions in Appendix B.

4.2 Annotation

From the annotation scheme, we develop a codebook with task documentation and illustrative examples. We create an annotation interface formatted as a spreadsheet of utterances to evaluate, preceding utterances for context, and checkboxes for each possible label. We recruit teachers with classroom experience at an elementary level as annotators, whose demographics are representative of the student population in the data. Demographics of annotators are reported in Appendix C. After training annotators via presentations, exercises, and discussions to apply labels according to the codebook, we ask annotators to participate in several annotation trials. Following each trial, we calculate inter-rater agreement scores to assess the consistency of labels across annotators. We address disagreements through feedback and discussion to clarify misunderstandings of task definitions, but we encourage annotators to

Annotation Task	Definition	Example	Label
Identifying language indicative of classroom management	Classroom management is defined by language unrelated to academic content that involves the management of the instructional environment, including establishing rules and procedures, maintaining student attention and engagement, behavior modification, counseling, and administration of supplies or group work.	Guys, real quick listen. Student M, please don't touch the stuff on my desk. I need you to sit and participate with your group. Eyes on me. One thing or two things that I really wanna stress.	Yes
		One pile on your desks, please. Put everything away now. Our math lesson is starting. Student A, will you collect them please?	Yes
		And how far is it, the whole distance? Four-fifths a mile. And she's already walked half. And you're trying to find how much further she has to go.	No
Identifying punitive and positive attitudes in classroom management language	Punitive attitudes are defined by prescriptive, pressuring, and inflexible language, and negative affect that seeks to shame, threaten, or otherwise punish students. Positive attitudes are defined by non-pressuring, informational language that values students' perspectives, nurtures students' growth, and maintains positive relationships with students.	Excuse me, if I need to, I will have you all put your heads down, and I will take that time off of recess. Turn your bodies and face forward. I do not appreciate this.	Punitive
		Do you understand what you need to do on your math paper? Good. Dry your eyes up. Take a break. And then you go ahead back because you only got about five more minutes. Do you need to go get a drink of water or something? That might help.	Positive
		So let us proceed. On your desk you have white sheets of paper. I want you to each take one. Everybody, just one white sheet, and you need a pencil.	Neutral

Table 1: Examples from our annotated data for each defined annotation task, with majority assigned labels.

give ratings subjectively based on their interpreted tone of the utterance text, influenced by their classroom experience.

The annotation process involves two phases. In the first phase, three teachers label whether utterances meet the definition of containing classroom management language. We select 17,523 teacher utterances from the dataset via stratified random sampling to ensure an even distribution of observations across CLASS observation scores and classroom student demographics. We define six strata from a principle component analysis over scores from the classroom organization and emotional support dimensions of the CLASS observation protocol and four strata from a principle component analysis over variables about the classroom racial composition of students. We perform two annotation trials, each including 200 utterances selected via simple random sampling, and obtain a high average inter-rate agreement (Fleiss κ = .812). We then assign the 17,523 utterances randomly to each annotator, resulting in a single label for each utterance.

In the second phase, six teachers label punitive, positive, and neutral attitudes in the

1,422 utterances labeled in the previous phase as containing classroom management language. We perform four annotation trials, each including 50 utterances selected via simple random sampling, and obtain an average inter-rater agreement of Fleiss κ = 0.364. Our inter-rater agreement values obtained in widely-used classroom observation protocols such as MQI and CLASS (Pianta et al., 2008). The lower agreement value for classifying attitudes can be explained in part by varying reliance on the "neutral" category, as the percentage agreement for "punitive" and "positive" ratings are 0.748 and 0.796, respectively. Additionally, rater disagreement is expected due to the subjective nature of interpreting tone and intentions. We assigned each of the 1,422 utterances randomly to two raters. To obtain a single label for each utterance, we first convert the labels to numeric values (punitive = -1, neutral = 0, positive = 1), then z-score labels within each rater to account for between-rater differences, and lastly average z-scores for each example.

4.3 Model Development

Using data we collected from each annotation task, we train machine learning models to predict language-based measures of classroom management practice. We employ standard techniques for text classification by fine-tuning pre-trained language models. These models, such as BERT (Devlin et al., 2018) and RoBERTa (Y. Liu et al., 2019), initially learn word representations from vast amounts of text data. Fine-tuning involves further training these models on our 17,523 utterances annotated for the presence of classroom management language and 1,422 utterances annotated for the representation of punitive and positive attitudes. In doing so, we adapt them to perform our target prediction tasks. The resulting fine-tuned models predict the following dimensions for each new utterance:

- **CM:** A binary value indicating whether or not the utterance contains classroom management language.
- **Punitiveness:** A continuous value indicating the degree to which an utterance reflects

a punitive or positive attitude. To increase interpretability, we flip and scale the raw outputs (between -1 to 1, as described in Section 4.2) to range between 0 and 1, where larger values represent more punitive attitudes.

We validate the performance of our models through the traditional machine-learning paradigm of cross-validation. This process involves dividing the annotated data into subsets for training and evaluation and obtaining average metrics for the accuracy of model predictions across multiple rounds of testing. Our best model for identifying classroom management language achieves an accuracy of 0.933, and our best model for predicting punitiveness achieves a strong positive Spearman correlation ρ of 0.577 ($p < 0.001$) with human expert labels. Further details on our data pre-processing, modeling approach, parameters, and performance metrics are available in Appendix D.

4.4 Predicting Values for the Entire Dataset

We apply our models to the full set of 295,709 teacher utterances in the NCTE dataset, using our models to predict CM and Punitiveness for each utterance. With these outputs, we calculate the rate and punitiveness of classroom management for each observation transcript. We define the rate of classroom management as the number of teacher utterances containing classroom management language per hour of classroom observation. We define the punitiveness of classroom management as the average punitiveness score across classroom management utterances. For analyses that use teacher- or student-level dependent variables, we mean aggregate transcript-level values to the teacher- and student-level, respectively.

Statistics of computed metrics. Across transcripts in the dataset, we observe a mean rate of 20 (SD = 15) classroom management utterances per hour, representing on average 7% of teacher utterances per hour. Rates of CM varied greatly—while 30 observations included 0 classroom management utterances, 20% or more of teacher utterances were related to classroom management in 34 observations. Since punitiveness ratings were standardized

before training the model, the computed scores are close to the midpoint of the 0-1 range (M=0.481, SD = 0.079). By estimating a variance decomposition model, we find that differences between teachers accounts for 26% of the variance in automated measures on average, while differences between schools accounts for only 7% (see Appendix G). The rate and punitiveness of classroom management utterances have a strong positive Spearman correlation ρ of 0.218 ($p < 0.001$).

Words associated with punitiveness. To make concrete the abstract characterization of attitudes, we present words and phrases from the teacher utterances that represent the measured dimensions. We explain our method for extracting these examples in Appendix F. Language from the most punitive utterances commands student bodies (*sit down* and *sit up*) and student expression (*shh*, *talking*, and *yelling*). These examples include negative descriptors like *disrespectful*, *unacceptable*, *distracting*, and *rude*. Punitive language focuses on what students can not do (*stop* and *you cannot*) compared with more positive utterances (*you can* and *go ahead*). Language from the most positive utterances involves words that personally address students, such as *you*, *honey*, and *sweetie*, and terms of appreciation and praise.

5 RQ1: What Does CM Language Predict?

We evaluate the extent to which the frequency and punitivness of classroom management language predicts observation scores of classroom management competencies, teacher and student perceptions of classroom climate, and students' test scores. These analyses produce empirical evidence to inform our understanding of effective classroom management by relating teacher behavior to observed competencies, classroom atmosphere and student learning.

5.1 Methods

We model indicators and outcomes of effective classroom management as a function of the rate and punitiveness of classroom management language and a series of covariates. Specifically, we estimate an ordinary least squares regression: $Y_i = \beta X_i + \theta Z_i + \epsilon_i$. We select the following indicators as dependent variables Y_i and describe each in further detail in Appendix A:

- **Observational scores:** We use scores for each item within the classroom organization and emotional support categories of the CLASS observational protocol (Pianta et al., 2008). We additionally use the holistic mathematical quality of instruction score (MQI5) derived from the MQI instrument (Hill et al., 2008), as well as the dimension of MQI assessing the extent to which classroom work is connected to mathematics (CWCM).
- **Survey responses:** We use teacher survey responses on items related to the frequency of behavior management, loss of instructional time, and perceptions of disrespect. We use student survey responses on items related to perceptions of their own behavior.
- **Student outcomes:** We use student end-of-the-year standardized exam scores in math.

The predictor X_i indicates our automated measures of the rate and punitiveness of classroom management language, which we regress separately, and β is our parameter of interest. The covariates represented in Z_i include teacher self-reported demographics (male, Black, Hispanic, and years of experience), classroom demographics based on administrative data (proportion of male, Black, and Hispanic students, proportion of students qualifying for free or reduced-price lunch, and proportion of students with special education or English language learner status), classroom features (grade level and class size), as well as district and school year fixed effects. The regressions correlating automated measures against standardized test scores additionally control for standardized student test scores from the previous academic

year. When using punitiveness as the dependent variable, we filter the dataset to only those observations with at least one classroom management utterance and additionally control for the number of classroom management utterances per transcript. We cluster standard errors at the teacher level.

5.2 Results

We present results from these regressions in Table 2. We find significant correlations between instructional outcomes and the rate and punitiveness of CM language. In Panel A, we report estimates for CLASS scoring items in the classroom organization and emotional support categories. The rate of classroom management is strongly correlated with CLASS ratings for behavior management, productivity, student engagement, negative climate, and teacher sensitivity. An increase in CM rate by 10 classroom management utterances per hour corresponds with a 0.19 ($p < 0.001$) standard deviations decrease in behavior management score, a 0.15 ($p < 0.001$) standard deviations decrease in productivity score, a 0.07 ($p < 0.001$) standard deviations decrease in the student engagement score, and a 0.04 ($p < 0.001$) standard deviations decrease in teacher sensitivity score. The rate of classroom management is also significantly correlated with the CWCM MQI dimension but displays no significant relationship with the overall MQI score.

The punitiveness of classroom management is strongly correlated with CLASS ratings for behavior management, positive climate, negative climate, teacher sensitivity, and regard for student perspectives. One standard deviation increase in the punitiveness of classroom management utterances corresponds with a 0.217 ($p < 0.001$) standard deviations decrease in the behavior management score, a 0.258 ($p < 0.001$) standard deviations decrease in the positive climate score, a 0.250 ($p < 0.001$) standard deviations decrease in the teacher sensitivity score, and a 0.212 ($p < 0.001$) standard deviations decrease in observation scores for regard for student perspectives. We additionally observe a significant correlation between the punitiveness of classroom management and overall MQI score.

Dependent Variable	CM Rate	n	Punitiveness	n
Panel A: Observational Scores				
CLASS: Behavior Management	-0.019***	1570	-0.217***	1540
	(0.002)		(0.033)	
CLASS: Productivity	-0.015***	1570	-0.103**	1540
	(0.002)		(0.034)	
CLASS: Student Engagement	-0.007***	1570	-0.176***	1540
	(0.002)		(0.031)	
CLASS: Positive Climate	-0.004*	1570	-0.258***	1540
	(0.002)		(0.035)	
CLASS: Negative Climate	0.013***	1570	0.236***	1540
	(0.002)		(0.039)	
CLASS: Teacher Sensitivity	-0.006***	1570	-0.250***	1540
	(0.002)		(0.034)	
CLASS: Regard for Student Perspectives	-0.002	1570	-0.212***	1540
	(0.002)		(0.035)	
MQI5	-0.003	1570	-0.124***	1543
	(0.002)		(0.033)	
MQI: Classroom Work is Connected to Mathematics	-0.012***	1570	0.031	1543
	(0.002)		(0.029)	
Panel B: Survey Responses				
Teacher: Frequency of reprimanding students	0.012***	536	0.273***	533
	(0.003)		(0.084)	
Teacher: Frequency of losing time to student misbehavior	0.010**	537	0.147**	534
	(0.004)		(0.075)	
Teacher: Frequency of feeling disrespected	0.011**	537	0.229*	534
	(0.004)		(0.073)	
Student: My behavior in this class is good	-0.004***	10980	-0.096***	10919
	(0.001)		(0.020)	
Student: My behavior is a problem for the teacher in this class	0.003***	10810	0.094***	10753
	(0.001)		(0.022)	
Panel C: Student Outcomes				
State Standardized Exam Score in Math	-0.005***	10741	-0.010	10677
	(0.001)		(0.019)	

Table 2: Correlations between the rate and punitiveness of classroom management language, observation scores, survey responses and student outcomes. Each value displays the results from a separate regression. All models include district, school year, and grade level fixed effects, as well as teacher and class demographic covariates. All estimates are in standard deviation units with respect to the dependent variable. CM Rate refers to the number of classroom management utterances per hour of observed class time. Punitiveness is the scaled attitude score in standard deviation units. Robust standard errors clustered at the teacher level are in parentheses. $*p < .05, **p < .01, ***p < .001$.

Estimates for survey items are reported in Panel B. The rate and punitiveness of classroom management are both strongly associated with teacher-reported frequencies of reprimanding students and losing time to student misbehavior. Teacher survey items reporting of the frequency of feeling disrespected is more significantly correlated with the classroom management rate than with punitiveness. The rate and punitiveness of classroom management are both strongly correlated with student perceptions of whether their behavior in class is a problem for the teacher. Finally, we report estimates for student outcomes in Panel C. We find a significant correlation between student exam scores and classroom management rate, but no significant relationship with punitiveness. An increase of 10 classroom management utterances per hour is associated with a 0.05 ($p < 0.001$) standard deviation decrease in exam scores, controlling for prior academic performance.

The systematic and strong correlations observed here indicate that the rate and punitiveness of classroom management language is a key predictor of effective instruction. The effect of language on observed classroom management competencies, classroom climate, and student learning are complicated, however, by contextual classroom factors and student and teacher identities. To understand the conditions that prompt different uses of classroom management language, we next analyze their relationship with teacher and classroom characteristics.

6 RQ2: What Predicts CM Language?

We examine the degree to which demographic and operational classroom features relate to the rate and punitiveness of classroom management language. By analyzing covariates as predictors of our automated measures, we identify potential drivers of disparities in classroom management practices.

6.1 Methods

We estimate ordinary least squares regressions, modeling automated measures as a function of classroom characteristics, including academic year and month of observation, grade level, class size, teacher experience, teacher demographics, and student classroom demographics. Unlike in Section 5, we now model CM Rate and Punitiveness as dependent variables. The same covariates are used in each regression, but at two different levels of aggregation. To examine the relationship between teacher characteristics and CM language, we aggregate data to the teacher level (Model 1). To examine relationships between the classroom characteristics and CM language, we perform analyses at the observation-level (Model 2). In each regression, we control for district and school-year fixed effects.

6.2 Results

We report estimates from these regressions in Table 3. We find that teacher demographic factors correlate with the rate of classroom management language. Male teachers use 4.65 ($p < 0.001$) fewer classroom management utterances per hour, indicating a 14% decrease compared to non-male teachers. These results corroborate studies finding that male teachers exhibit less controlling attitudes towards instructional management (Martin et al., 2006). Additionally, we find a negative but not-significant relationship between teacher experience level and the rate and punitiveness of classroom management. Though new teachers nationwide ask for support about managing classrooms (Freeman et al., 2014), new teachers may not be alone in requiring professional development in these competencies (Baker, 2005). While studies have found that teachers' years of experience affect their beliefs and attitudes toward classroom management, their behaviors in practice may result in similar aggregate levels of management language. More experienced teachers have been found to favor interventionist and controlling approaches (Martin & Shoho, 2000; Zafer & Aslihan, 2012) to instructional management while less experienced teachers apply control reactively in interpersonal management (Martin et al., 2006).

Independent Variable	CM Rate	n	Punitiveness	n
Panel A (Model 1): Teacher Demographic Covariates				
Years of Experience	-0.095	1573	-0.003	1543
	(0.089)		(0.005)	
Male	-4.647***	1573	0.097	1543
	(1.588)		(0.081)	
Black	-2.478	1573	0.022	1543
	(1.744)		(0.095)	
Hispanic	-4.259	1573	-0.146	1543
	(2.329)		(0.178)	
Panel B (Model 2): Observation Covariates				
Academic Year Month	0.563*	1573	0.037**	1543
	(0.292)		(0.014)	
Grade Level	-5.464***	1573	-0.048	1543
	(1.182)		(0.065)	
Class Size	0.293***	1573	0.002	1543
	(0.123)		(0.008)	
Panel C (Model 2): Student Demographic Covariates				
% Male	0.023	1573	-0.003	1543
	(0.053)		(0.002)	
% Black	0.112***	1573	0.008***	1543
	(0.037)		(0.002)	
% Hispanic	0.053	1573	0.004**	1543
	(0.041)		(0.002)	
% Free or Reduced Price Lunch	-0.037	1573	-0.002	1543
	(0.035)		(0.002)	
% Special Education Classification	0.059*	1573	-0.001	1543
	(0.041)		(0.002)	
% English Learner Classification	0.002*	1573	-0.001	1543
	(0.040)		(0.002)	

Table 3: Correlations between teacher and classroom covariates and our CM Rate and Punitiveness measures. The CM Rate and Punitiveness columns display results from separate regressions. All models include district and school-year fixed effects. Standard errors are in parentheses. $*p < .05, **p < .01, ***p < .001$.

We also find that classroom factors, such as class size and grade level, play a significant role in classroom management language. An increase of 1 student in a classroom is associated with 0.30 ($p < 0.001$) more classroom management utterances per hour. Teachers in 5th-grade classrooms use on average 5.50 ($p < 0.001$) fewer classroom management utterances per hour than those in 4th-grade classrooms (-24%). Additionally, the month of observation is significantly and positively correlated with both classroom management rate and punitiveness, indicating an increase in the use and punitiveness of classroom management practice over the course of the school year. This escalation reflects those found by Darling-Hammond et al. (2023) for rates of disciplinary incidents.

Notably, student racial demographics correlate significantly with the rate and punitiveness of teachers' classroom management language. A 10 percent increase in the proportion of Black students in a classroom is associated with 1.12 ($p < 0.001$) more classroom management utterances per hour and a 0.08 ($p < 0.001$) standard deviations increase in the punitiveness score. We also observe a weaker but significant correlation between the proportion of Hispanic students in a classroom and the punitiveness of classroom management utterances. A 10 percent increase in the proportion of Hispanic students is associated with a 0.04 ($p < 0.01$) standard deviations increase in the punitiveness score. These findings support a wider body of research detailing racial inequalities in the frequency of formalized disciplinary events and severity of punishments (J. Liu, Penner, & Gao, 2022; Barrett et al., 2021; J. Liu, Hayes, & Gershenson, 2022). We extend this literature by identifying these disparities at the level of classroom management discourse. The language used by teachers when managing their classrooms may play an important role in understanding later disparities in disciplinary action.

In these analyses, we establish the significant role of teacher and student demographics in predicting classroom management language. Our results position everyday classroom management interactions as a critical site of intervention for addressing racial disparities. They also raise questions about the nuanced effects of teacher experience on punitive attitudes

and management behaviors. We next contextualize these findings in the temporal dynamics of classrooms and map the evolution of teachers' attitudes in management language within observations.

7 RQ3: How Does the Use of CM Language Shift Within an Observation?

A unique advantage of language-based measures is their granularity, enabling evaluations of classroom dynamics at the utterance-level within observations. By observing changes in classroom management language over the duration of a single class period, we further examine the emergence of racial disparities and effect of teacher experience. These temporal analyses can provide a more nuanced understanding of the dynamic nature of classroom management practices. Knowing when and how the frequency and punitiveness of classroom management language escalates can help locate moments for intervention and facilitate professional learning.

7.1 Methods

To create time series data, we first divide the utterances in each transcript into ten equal-sized, sequential bins by word count. We divide observations by words instead of duration given the absence of exact timestamps in the transcript. If an utterance spans multiple bins, we assign it to the bin with the greater overlap. We mean-aggregate our CM Rate and Punitiveness measures within each of the ten bins. In order to describe the role of student race and teacher experience discussed in Section 6, we compare trends for subsets of observations. We isolate trends for classrooms in the top and bottom quartiles in terms of the proportion of Black students.

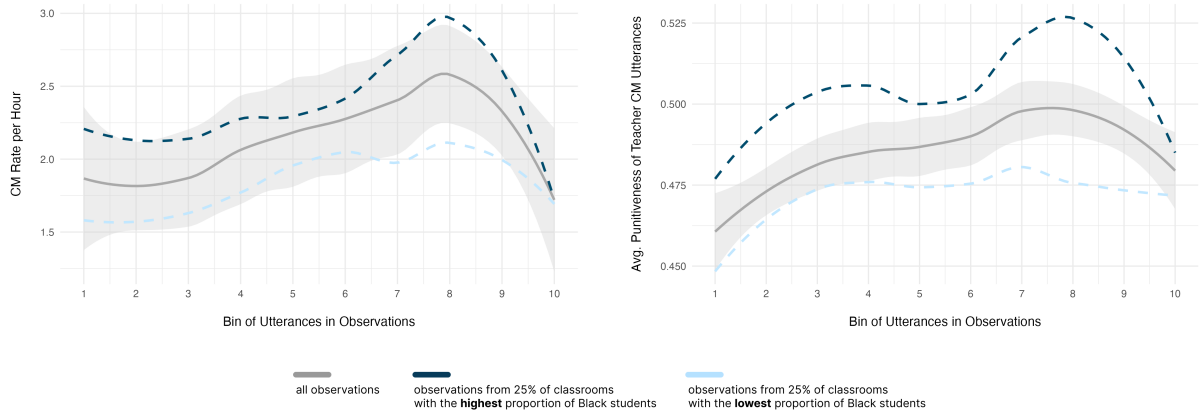

Figure 1: The variation in the rate and punitiveness of classroom management language over time within observations, calculated by dividing the teacher utterances in each transcript into ten sequential bins by word count. The dashed lines show comparative trends for the subsets of observations from classrooms with the highest and lowest proportion of Black students. The grey area indicates the 95% confidence interval for the overall trend.

7.2 Results

Figure 1 illustrates the trends in the rate and punitiveness of classroom management language within observations. Both measures exhibit an overall upward trend in the first eight bins of utterances, indicating that during this time teachers employ more classroom management utterances that also become progressively more punitive. In the last two bins of utterances, both measures decrease.

We also present isolated within-observation trends for the classrooms with the highest and lowest proportions of Black students. Racial disparities occur in the slopes of trajectories. The rate and punitiveness of of classroom management language escalate more severely for classrooms with the highest proportions of Black students. These escalations are most notable between the sixth and eighth bins of utterances for both measures. These findings extend previous literature about escalations of disciplinary events over the school year and findings of higher degrees of escalation for Black students (Darling-Hammond et al., 2023).

Situating our finding of patterns of escalation, research has suggested ways in which teacher–student relationships may worsen over time. Teachers who are inadequately pre-

pared to manage classrooms resort to reactive control and "survival skills" (Martin et al., 2006), escalating "minor" student misbehavior to levels with significant consequences Albin et al. (1995). Studies have also pointed to Black students' awareness of racial biases and subsequent loss of institutional trust as a factor in the escalation of racial disciplinary disparities (Okonofua, Walton, & Eberhardt, 2016; Yeager et al., 2017). Escalation may additionally signal factors such as the derailment of classroom routines and loss of student academic engagement. Though exploratory, our temporal analyses indicate sites of intervention for de-escalating the rate and punitiveness of classroom management language and reducing racial disparities.

8 Discussion

Our findings demonstrate the critical role of teachers' language in understanding and improving classroom management practices. By applying a natural language processing approach to transcripts of classroom dialogue, we contribute novel computational measures that identify classroom management language and punitive attitudes. The measures provide a novel lens through which we examine the outcomes, predictors, and dynamic nature of classroom management.

Our analyses reveal racial disparities in the frequency and punitiveness of classroom management language. Classrooms with higher proportions of Black students experience more and more punitive classroom management. These findings corroborate and extend previous literature showing racial disparities in school discipline (Girvan et al., 2021; Markowitz et al., 2023), highlighting potential in-class precursors to more extreme disciplinary practices. Given appropriate data, our measures can help further examine the link between in-class and out-of-class disciplinary actions, and the mechanisms underlying racial disparities.

Further, our results signal sites of intervention for improving classroom management practice. The rate and punitiveness of classroom management language escalate during

observations. These escalations occur more severely for classrooms with higher proportions of Black students, widening disparities. By identifying periods of escalation, we underscore the need for professional learning regarding de-escalation and identify opportunities for timing interventions to address classroom management spikes.

Finally, the systematic and significant relationships between our measures of classroom management language and observational measures of classroom management quality highlight the promise of low-cost automated tools to complement trained observers and facilitate classroom observation. The increasing prevalence of discourse data collected through traditionally taped classroom observations, teacher self-recording in professional development (Scornavacco et al., 2022), combined with developments in automated transcription (Radford et al., 2022), provide new opportunities to apply automated measures to diverse research settings. These measures may be applied as an early indicator in evaluating disciplinary reforms and as a benchmark to monitor teachers' classroom management training. Leveraging the affordances of computational methods, this study develops the infrastructure to move research into the classrooms of teachers experiencing the realities of classroom management, at scale.

8.1 Limitations and Future Directions

Our measures may face generalizability constraints due to the focus of the NCTE dataset on elementary math classrooms in underserved public schools. Questions of relevance are also valid, as the NCTE dataset was collected a decade ago, though research suggests that teaching practices have remained relatively constant over the past century (Cuban, 1993; D. K. Cohen & Mehta, 2017). Applying these measures to new settings and classroom transcripts may serve to test and improve the generalizability of the model while enabling analyses of variation between different types of instructional environments.

Additionally, our analysis is unable to explore the relationship between classroom management language and disciplinary outcomes for students. This limitation is partly due to

the high rate of missing in-school and out-of-school suspension values in the NCTE dataset and partly due to the infrequency of suspensions for the grade levels represented. Future applications of these measures should explore the missing link and test the hypothesis that the language of classroom management may predict downstream disciplinary action.

Further, though we find significant disparities in classroom management language by classroom racial demographics, we are unable to assess racial differences in the individual treatment of students. Because our data lacks mappings between individual student speakers and demographic features, we correlate classroom management practices with student demographics only at the classroom level. Future work should apply these measures to relational transcript data and investigate how teachers may use classroom management language differently depending on the identity of the students they address.

Moreover, future research can expand our investigation of the conditions that give rise to classroom management language and punitive attitudes via qualitative analyses of language, adding additional variables from the NCTE data (e.g. resources available to the teacher), and analyzing the indirect relationships between variables. As our temporal analysis indicated escalations in the rate and punitiveness of classroom management language, exploring the language at each stage can expose the evolution of topics and problems.

Finally, building upon these measures to improve teacher practices is the ultimate goal of this work. Future work can develop tools that enable self-led assessments of punitiveness or complement instructional coaching with AI-powered feedback to support teachers with practical training opportunities embedded in real-world classroom environments.

References

Albin, R. W., O'Brien, M., & Horner, R. H. (1995). Analysis of an escalating sequence of problem behaviors: A case study. *Research in Developmental Disabilities*, *16*(2), 133–147.

Anderson, L. M., Evertson, C. M., & Emmer, E. T. (1980). Dimensions in classroom management derived from recent research. *Journal of Curriculum Studies*, *12*(4), 343-356. Retrieved from https://doi.org/10.1080/0022027800120407 doi: 10.1080/0022027800120407

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

Baker, P. H. (2005). Managing student behavior: How ready are teachers to meet the challenge? *American secondary education*, 51–64.

Barrett, N., McEachin, A., Mills, J. N., & Valant, J. (2021). Disparities and discrimination in student discipline by race and family income. *Journal of Human Resources*, *56*(3), 711–748.

Bear, G. G., Slaughter, J. C., Mantz, L. S., & Farley-Ripple, E. (2017). Rewards, praise, and punitive consequences: Relations with intrinsic and extrinsic motivation. *Teaching and Teacher Education*, *65*, 10–20.

Brophy, J. (2006). History of research on classroom management. *Handbook of classroom management: Research, practice, and contemporary issues*, 17–43.

Buckmaster, D. (2016). From the eradication of tolerance to the restoration of school community: Exploring restorative practices as a reform framework for ethical school discipline. *Values and Ethics in Educational Administration*, *12*(3), n3.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311-337. Retrieved from https://doi.org/10.1177/0013164414539163 doi: 10.1177/0013164414539163

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., & Lovison, V. S. (2020). Bias in the air: A nationwide exploration of teachers' implicit racial attitudes, aggregate bias, and student outcomes. *Educational Researcher*, *49*(8), 566–578.

Cohen, D. K., & Mehta, J. D. (2017). Why reform sometimes succeeds: Understanding the conditions that produce reforms that last. *American Educational Research Journal*, *54*(4), 644–690.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387.

Cuban, L. (1993). *How teachers taught: Constancy and change in american classrooms, 1890-1990*. Teachers College Press.

Cummings, C. B. (2000). *Winning strategies for classroom management*. ASCD.

Darling-Hammond, S., Ruiz, M., Eberhardt, J. L., & Okonofua, J. A. (2023). The dynamic nature of student discipline and discipline disparities. *Proceedings of the National Academy of Sciences*, *120*(17), e2120417120.

Demszky, D., & Hill, H. (2022). The NCTE transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873*.

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

Evertson, C. M., Weinstein, C. S., et al. (2006). Classroom management as a field of inquiry. *Handbook of classroom management: Research, practice, and contemporary issues*, *3*(1), 16.

Floress, M. T., Beschta, S. L., Meyer, K. L., & Reinke, W. M. (2017). Praise research trends and future directions: Characteristics and teacher training. *Behavioral Disorders*, *43*(1), 227–243.

Freeman, J., Simonsen, B., Briere, D. E., & MacSuga-Gage, A. S. (2014). Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs. *Teacher Education and Special Education*, *37*(2), 106–120.

Gage, N. A., Scott, T., Hirn, R., & MacSuga-Gage, A. S. (2018). The relationship between teachers' implementation of classroom management practices and student behavior in elementary school. *Behavioral disorders*, *43*(2), 302–315.

Gettinger, M., & Walter, M. J. (2012). Classroom strategies to enhance academic engaged time. In *Handbook of research on student engagement* (pp. 653–673). Springer.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. rel 2017-191. *Regional Educational Laboratory Mid-Atlantic*.

Girvan, E. J., McIntosh, K., & Santiago-Rosario, M. R. (2021). Associations between community-level racial biases, office discipline referrals, and out-of-school suspensions. *School Psychology Review*, *50*(2-3), 288–302.

Gregory, A., Skiba, R. J., & Noguera, P. A. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational researcher*, *39*(1), 59–68.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, *26*(4), 430–511.

Hunkins, N., Kelly, S., & D'Mello, S. (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* (pp. 230–238).

Ingersoll, R. M., & Smith, T. M. (2003). The wrong solution to the teacher shortage. *Educational leadership*, *60*(8), 30–33.

Kane, T., Hill, H., & Staiger, D. (2015). *National center for teacher effectiveness main study. icpsr36095-v2.* Inter-university Consortium for Political and Social Research (distributor ....

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464.

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*(3), 643–680.

Landrum, T. J., & Kauffman, J. M. (2013). Behavioral approaches to classroom management. In *Handbook of classroom management* (pp. 57–82). Routledge.

Lekwa, A. J., Reddy, L. A., & Shernoff, E. S. (2019). Measuring teacher practices and student academic engagement: A convergent validity study. *School Psychology*, *34*(1), 109.

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

Liu, J., Hayes, M. S., & Gershenson, S. (2022). Jue insight: From referrals to suspensions: New evidence on racial disparities in exclusionary discipline. *Journal of Urban Economics*, 103453.

Liu, J., Penner, E. K., & Gao, W. (2022). Troublemakers? the role of frequent teacher referrers in expanding racial disciplinary disproportionalities. *EdWorkingPapers. com*.

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

Markowitz, D. M., Kittelman, A., Girvan, E. J., Santiago-Rosario, M. R., & McIntosh, K. (2023). Taking note of our biases: How language patterns reveal bias underlying the use of office discipline referrals in exclusionary discipline. *Educational Researcher*, 0013189X231189444.

Martin, N. K., & Shoho, A. R. (2000). Teacher experience, training, & age: The influence of teacher characteristics on classroom management style.

Martin, N. K., Yin, Z., & Mayall, H. (2006). Classroom management training, teaching experience and gender: Do these variables impact teachers' attitudes and beliefs toward classroom management style?. *Online Submission*.

McLeod, J., Fisher, J., & Hoover, G. (2003). *The key elements of classroom management: Managing time and space, student behavior, and instructional strategies*. ASCD.

Mitchell, M. M., & Bradshaw, C. P. (2013). Examining classroom influences on student perceptions of school climate: The role of classroom management and exclusionary discipline strategies. *Journal of school psychology*, *51*(5), 599–610.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.

Okonofua, J. A., Paunesku, D., & Walton, G. M. (2016). Brief intervention to encourage empathic discipline cuts suspension rates in half among adolescents. *Proceedings of the National Academy of Sciences*, *113*(19), 5221–5226.

Okonofua, J. A., Walton, G. M., & Eberhardt, J. L. (2016). A vicious cycle: A social–psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science*, *11*(3), 381–398.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual k-3.* Paul H Brookes Publishing.

Powell, R., Cantrell, S. C., Malo-Juvera, V., & Correll, P. (2016). Operationalizing culturally responsive instruction: Preliminary findings of criop research. *Teachers College Record*, *118*(1), 1–46.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.*

Rajapakse, T. (2019). Simple transformers. *URL: https://simpletransformers. ai/[accessed 2022-08-25]*.

Reeve, J. (2016). Autonomy-supportive teaching: What it is, how to do it. *Building autonomous learners: Perspectives from research and practice using self-determination theory*, 129–152.

Reinke, W. M., Herman, K. C., & Sprick, R. (2011). *Motivational interviewing for effective classroom management: The classroom check-up.* Guilford press.

Reinke, W. M., Herman, K. C., & Stormont, M. (2013). Classroom-level positive behavior supports in schools implementing sw-pbis: Identifying areas for enhancement. *Journal of Positive Behavior Interventions*, *15*(1), 39-50. Retrieved from `https://doi.org/10.1177/1098300712459079` doi: 10.1177/1098300712459079

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., & Goel, N. (2011). Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers. *School psychology quarterly*, *26*(1), 1.

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., ... Graesser, A. (2014). Domain independent assessment of dialogic properties of classroom discourse. *Grantee Submission*.

Scornavacco, K., Jacobs, J., & Clevenger, C. (2022). Automated feedback on discourse moves: Teachers' perceived utility of a big data tool. In *Annual meeting of the american educational research association.*

Shinn, M. R., Ramsey, E., Walker, H. M., Stieber, S., & O'Neill, R. E. (1987). Antisocial behavior in school settings: Initial differences in an at risk and normal population. *The Journal of Special Education*, *21*(2), 69–84.

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and treatment of children*, 351–380.

Skiba, R. J., Horner, R. H., Chung, C.-G., Rausch, M. K., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of african american and latino disproportionality in school discipline. *School psychology review*, *40*(1), 85–107.

Smith, D., Fisher, D., & Frey, N. (2015). *Better than carrots or sticks: Restorative practices for positive classroom management*. ASCD.

Stone, C., Donnelly, P. J., Dale, M., Capello, S., Kelly, S., Godley, A., & D'Mello, S. K. (2019). Utterance-level modeling of indicators of engaging classroom discourse. *International Educational Data Mining Society*.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? a cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of teacher Education*, *62*(4), 339–355.

Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., & Sumner, T. (2021). Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. *arXiv preprint arXiv:2105.07949*.

Sutherland, K. S., Alder, N., & Gunter, P. L. (2003). The effect of varying rates of opportunities to respond to academic requests on the classroom behavior of students with ebd. *Journal of Emotional and behavioral Disorders*, *11*(4), 239–248.

Van Acker, R., Grant, S. H., & Henry, D. (1996). Teacher and student behavior as a function of risk for aggression. *Education and Treatment of Children*, 316–334.

Wallace, T. L., Parr, A. K., & Correnti, R. J. (2020). Assessing teachers' classroom management competency: A case study of the classroom assessment scoring system–secondary. *Journal of psychoeducational assessment*, *38*(4), 475–492.

Wallace, T. L., Sung, H. C., & Williams, J. D. (2014). The defining features of teacher talk within autonomy-supportive classroom management. *Teaching and Teacher Education*, *42*, 34–46.

Weinstein, C. S., Tomlinson-Clarke, S., & Curran, M. (2004). Toward a conception of culturally responsive classroom management. *Journal of teacher education*, *55*(1), 25–38.

Yeager, D. S., Purdie-Vaughns, V., Hooper, S. Y., & Cohen, G. L. (2017). Loss of institutional trust among racial and ethnic minority adolescents: A consequence of procedural injustice and a cause of life-span outcomes. *Child development*, *88*(2), 658–676.

Zafer, Ü., & Aslihan, Ü. (2012). The impact of years of teaching experience on the classroom management approaches of elementary school teachers. *International journal of Instruction*, *5*(2).

A NCTE Dataset Variables

We describe all variables from the NCTE dataset used in our regression analyses in Table 4. We report the percentage of missing values and specify pre-processing aggregations. Mean and standard deviation values are included.

B Additional Measures: Praise, Opportunities for Response, and Explanatory Rationale

Our annotation schema included three additional dimensions for pertinent talk moves associated with classroom management literature. We define the annotation tasks of identifying praise, invitations for student perspectives, and provisions of explanatory rationale. These talk moves are recommended in effective and autonomy-supportive classroom management practice. Effective management is linguistically characterized by giving specific praise contingent on appropriate behavior (Floress et al., 2017; Simonsen et al., 2008), a positive ratio of supportive to corrective language, and providing students with opportunities to respond (Sutherland et al., 2003). Autonomy-supportive classroom management interactions practice transparency, providing explanatory rationales for requests and procedures, and invite student expression (Reeve, 2016; Wallace et al., 2014). An utterance contains praise if it includes encouraging, thanking, or complimentary language, seeking to provide positive feedback and reinforcement for student behavior. An utterance invites student perspectives if it contains questions asking for student input or assessing student engagement, or language seeking to understand students' needs. An utterance provides explanatory rationales if it contains language explaining the reasoning behind requests, rules, and procedures, seeking to help students understand expected behavior. Table 5 details definitions and examples for each dimension of this annotation schema.

Annotators labeled the occurrence of praise, opportunities for student response, and explanatory rationale in classroom management utterances in the second annotation phase. We obtained average inter-rater agreement values of Fleiss κ = 0.823 for identifying praise, κ = 0.542 for identifying opportunities for student response, κ = 0.483 for identifying explanatory rationale. Across the two raters assigned to each utterance, we consolidated responses based on whether either rater indicated the presence of the talk move in the utterance.

C Annotator Demographics

We collected annotator demographics from a survey. Five annotators identified as female, three annotators identified as male, and one annotator identified as non-binary. Two annotators identified as White or Caucasian, four annotators identified as Asian, two annotators identified as Black or African American, and one annotator identified as Hispanic or Latinx. Two annotators reported one to two years of experience, two annotators reported three to

Variable	Description	% Missing	M	Std.
Observation Scores				
CLASS: Behavior Management	Classroom organization dimensions of the Classroom Assessment Scoring System (CLASS) observational instrument. Scores range from 1 (low) to 7 (high). Scored in 15-minute segments and mean aggregated at the observation level.	0.2	6.10	0.83
CLASS: Productivity		0.2	6.36	0.73
CLASS: Student Engagement		0.2	5.25	0.89
CLASS: Positive Climate	Emotional support dimensions of the Classroom Assessment Scoring System (CLASS) observational instrument. Scores range from 1 (low) to 7 (high). Scored in 15-minute segments and mean aggregated at the observation level.	0.2	4.67	0.10
CLASS: Negative Climate		0.2	1.18	0.41
CLASS: Teacher Sensitivity		0.2	4.61	0.82
CLASS: Regard for Student Perspectives		0.2	3.53	0.96
MQI: Classroom Work is Connected to Mathematics	Dimension of the Mathematical Quality of Instruction (MQI) observational instrument for whether classroom work is connected to mathematics. Values are 0 (no) and 1 (yes), scored in 7.5-minute intervals and mean aggregated at the observation level.	0.0	0.95	0.09
MQI5	Dimension of the Mathematical quality of Instruction (MQI) for the quality of the whole-lesson, scored at the observation level from 1 (low) to 5 (high).	0.0	2.96	0.69
Classroom Characteristics				
Class Size	Number of students in the classroom at the time of observation.	1.8	20.9	5.08
Grade Level	Grade level of the student, aggregated by class identifier at the observation level.	2.4	4.47	0.49
% Male	Binary indicator for whether the student is male, aggregated by class identifier at the observation level.	1.8	0.17	0.13
% Black	Binary indicator for whether the student is Black, aggregated by class identifier at the observation level.	1.8	0.42	0.26
% Hispanic	Binary indicator for whether the student is Hispanic, aggregated by class identifier at the observation level.	1.8	0.24	0.23
% Free or Reduced Price Lunch	Binary indicator for whether the student receives free or reduced lunch in the year of observation, aggregated by class identifier at the observation level.	1.8	0.67	0.26
% Special Education Classification	Binary indicator for whether the student has special education status in the year of observation, aggregated by class identifier at the observation level.	1.8	0.14	0.17
% English Learner Classification	Binary indicator for whether the student has limited English proficiency in the year of observation, aggregated by class identifier at the observation level.	1.8	0.22	0.25
Teacher Surveys				
Frequency of reprimanding students	Teacher survey item for the frequency of reprimanding students while teaching math. Values range from 1 (rarely or never) to 5 (always).	2.1	2.13	0.97
Frequency of losing time to student misbehavior	Teacher survey item for the frequency of losing time to student misbehavior while teaching math. Values range from 1 (rarely or never) to 5 (always).	2.1	1.86	0.99
Frequency of feeling disrespected	Teacher survey item for the frequency of feeling disrespected while teaching math. Values range from 1 (rarely or never) to 5 (always).	2.1	1.46	0.80
Teacher Characteristics				
Years of Experience	Teacher's years of experience, including school year of survey.	0.8	10.55	6.86
Male	Binary indicator for whether the teacher is male.	0	0.17	0.38
Black	Binary indicator for whether the teacher is Black.	0	0.19	0.39
Hispanic	Binary indicator for whether the teacher is Hispanic.	0	0.03	0.18
Student Surveys				
My behavior in this class is good	Student survey item for student agreement with the sentiment that their behavior in class is good. Values range from 1 (totally untrue) to 5 (totally true).	6.6	4.23	0.89
My behavior is a problem for the teacher in this class	Student survey item for student agreement with the sentiment that their behavior in class is a problem for the teacher. Values range from 1 (totally untrue) to 5 (totally true).	8.1	1.75	1.15
Student Outcomes				
State Standardized Exam Score in Math	Student state math test score in year of observation, standardized.	3.2	0.05	0.93
State Standardized Exam Score in Math, Prior Year	Student state math test score in year prior to year of observation, standardized.	9.4	0.06	0.93

Table 4: Variables used in regression analyses, including percentage of missing values, mean, and standard deviation values.

Annotation Task	Definition	Example	Label
PR: Identifying praise in classroom management language	Praise is defined by encouraging, thanking, or complimentary language, seeking to provide positive feedback and reinforcement for student behavior.	All right, I like how you guys are working together and trying to figure this out. Let's see; did we figure it out?	Yes
		Student M, I like the way you're paying attention and you're being cooperative.	Yes
SR: Identifying opportunities for student response in classroom management language	Opportunities for student response are defined by questions that ask for student input or assess student engagement or readiness, and language seeking to understand students' needs, wants, preferences, priorities, goals, and emotions.	All right. Everyone has a pencil and is ready to go?	Yes
		Student R, how did you do, baby? You feeling okay today?	Yes
ER: Identifying provision of explanatory rationale in classroom management language	Explanatory rationale is defined by language explaining the reasoning behind requests, rules, and procedures, seeking to help students understand expected behavior.	So I only heard a little bit of what you said, but I'm wondering if you would be willing to come and explain your thinking up here, because what you were saying was really great, but I could only hear a little bit, okay?	Yes
		You're going to have to be quieter than you are right now. Let's move the table down a little bit, and then we'll move it back at the end. [Moving table] Ssssh, because you have Miss H's class there and the other class is there, so you can't be so loud.	Yes

Table 5: Examples from our annotated data for each defined annotation task, with majority assigned labels.

four years of experience, two annotators reported five to six years of experience, and three annotators reported more than eight years of experience.

D Modeling and Validation Procedures

We follow the state-of-the-art approach for performing text classification, which involves fine-tuning pre-trained language models. Pre-trained language models, such as BERT (Devlin et al., 2018) and RoBERTa (Y. Liu et al., 2019), are initially trained on massive amounts of unlabeled text data and can learn rich vector representations of words based on word co-occurrence patterns in the data. These representations can be fine-tuned by further training the models using a smaller task-specific dataset. In doing so, the models can be adapted to make predictions according to the requirements of a specific application domain. This procedure has been shown to perform well in classifying talk moves in classroom discourse (Suresh et al., 2021; Demszky & Hill, 2022). Using our annotated data, we fine-tuned BERT and RoBERTa (Y. Liu et al., 2019), selecting the one with better performance for each measure.

For the 17,523 utterances annotated for containing classroom management language, we fine-tuned a BERT-base binary classification model (Devlin et al., 2018) to predict for each new utterance one of two possible labels: whether or not it contains classroom management language. For the 1,422 utterances annotated for representing punitive and positive attitudes, we first converted the categorical labels to a numeric scale of punitiveness. Punitive utterances were converted to 1, neutral utterances were converted to 0, and positive utterances were converted to -1. To account for between-rater differences, we calculated z-scores for each rater's annotations and averaged z-scores across the two raters who were assigned to rate each utterance. We fine-tuned a RoBERTa-base regression model to predict for each

new utterance a punitiveness score Y. Liu et al. (2019).

We trained our models using the Simple Transformers library (Rajapakse, 2019). We specify 5 training epochs and a batch size of 8. For our binary classification models, we additionally specify a weight parameter. Because utterances containing classroom management language represented only 10% of the annotated dataset, the number of examples for each label is unbalanced. A commonly used tactic to deal with imbalanced datasets to improve prediction accuracy is to assign weights to each label.

We validated our models through 5-fold cross-validation. In this traditional machine learning paradigm, the annotated data is evenly divided into 5 subsets. The model is trained using data from 4 subsets and then evaluated, using standard machine learning metrics, on the remaining held-out test set. We perform this process 5 times until each subset is held out once for evaluation, then average the performance metrics from each test set. Our best model for identifying classroom management language achieves an accuracy of 0.933, precision of 0.571, F1 of 0.602, and recall of 0.671. Our best model for predicting punitiveness achieves a strong positive Spearman correlation ρ of 0.577 ($p < 0.001$) with human expert labels.

E Distribution of Measures

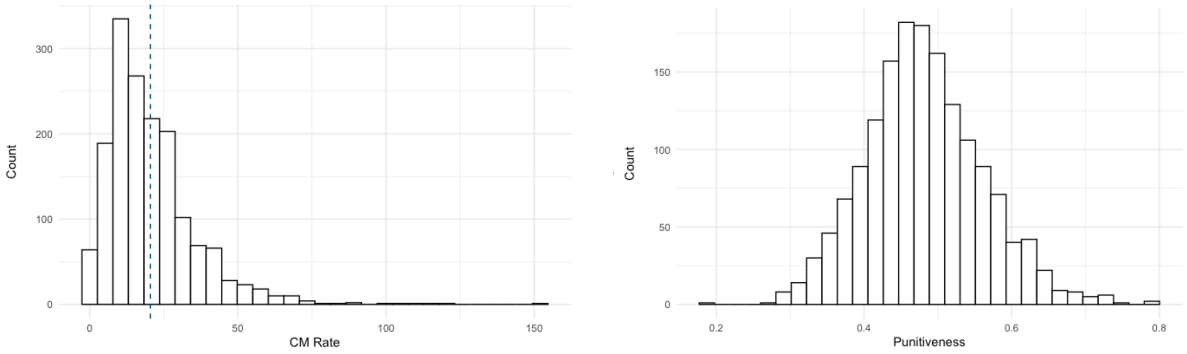

Figure 2: Distributions of CM Rate and Punitiveness measures in the dataset of 295,709 teacher utterances. The mean CM Rate is 20 and mean Punitiveness is a standardized value.

F Lexical Analysis

We examine the linguistic properties of classroom management language by analyzing the lexical differences between punitive and positive attitudes. Due to the affective nature of this dimension, there is limited articulation of teacher language choices associated with attitudes. By investigating how punitiveness manifests linguistically, we aim to characterize the language of positive interactions.

We first construct a vocabulary of single words (unigrams) and two-word phrases (bigrams) obtained by counting words in the dataset, after lowercasing, removing apostrophes and separating terms at spaces or punctuation marks. These unigrams and bigrams, collectively referred to as tokens, yield a vocabulary of 199,215 unique tokens.

We quantify the extent to which each token is associated with punitiveness by computing their z-scored log odds ratio via the method described in section 3.4 of (Monroe et al., 2008). This method has been designed to identify words that distinguish two groups of texts — in our case, the most and least punitive utterances — while controlling for the baseline frequency of words in the given dataset — in our case, the full set of classroom management utterances. The method produces a score for each token, which represents, in units of standard deviation, where the token lies on the spectrum of being associated with one vs the other group of texts. Tokens associated with the least punitive language indicated by negative values while words associated with the most punitive language are indicated by positive values.

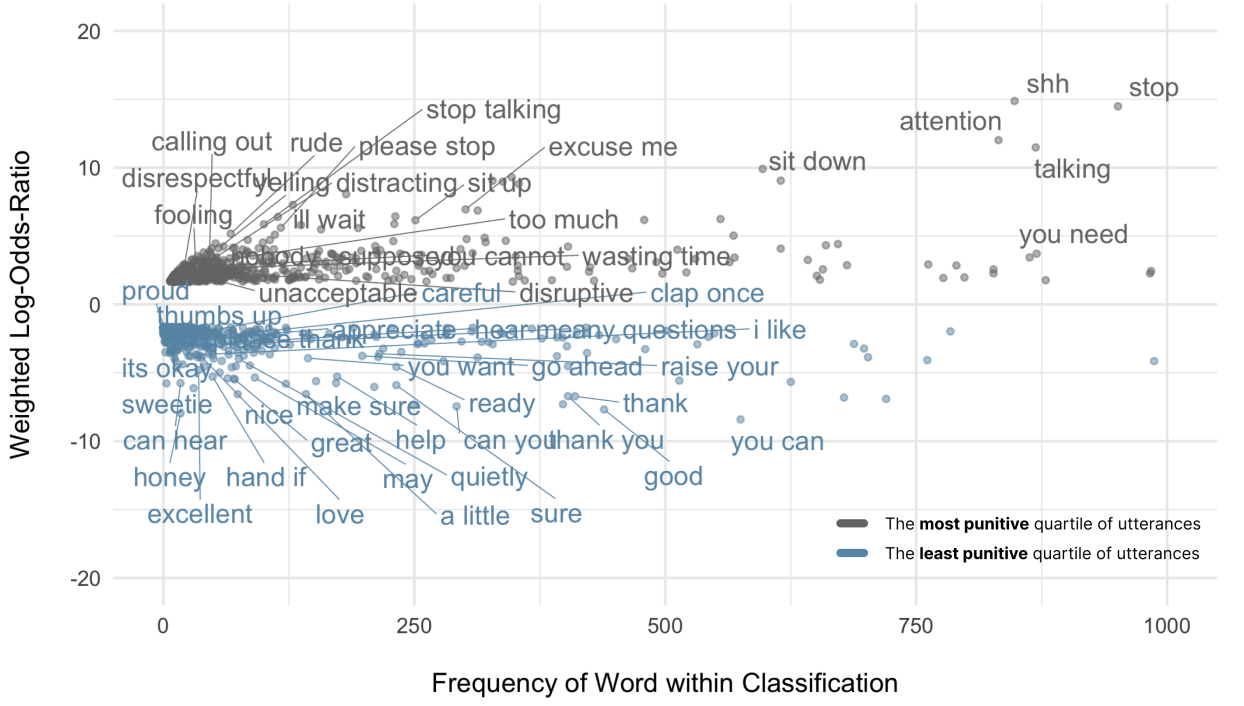

Figure 3: Top words and phrases from the most and least punitive utterances.

To create two groups of texts, we take the 6,263 classroom management utterances in the top quartile of punitiveness scores and the 7,061 classroom management utterances in the bottom quartile of punitiveness scores. For each token in our vocabulary, we compute its z-scored log-odds-ratio of belonging to either of the two sets while using the full set of 20,176 classroom management utterances to compute baseline frequencies. We filter these estimates to retain only critical z-score values when using a 95 percent confidence level, where the absolute values of z-scores exceed 1.96. This process allows us to identify the most polarized tokens, which exhibit significant associations with either the most punitive or the least punitive language.

Figure 3 shows the tokens most significantly associated with punitive vs positive teacher utterances, plotted against their frequencies within their respective categories.

G Variance Decomposition

We run a cross-classified multilevel model that decomposes the variance of each measure of classroom management language into teacher, school, and residual error component factors. The proportion of variance attributed to the teacher is reported in table 6.

Variable	Teacher	School	Residual
CM Rate	37.19	5.43	57.38
Punitiveness	26.11	7.19	66.7

Table 6: Variance components of teacher classroom management language measures.

H Temporal Dynamics by Observation Month

We plotted the mean-aggregated rate and punitiveness of classroom management language across observation months. Figure 4 represents the trends in the rate and punitiveness of

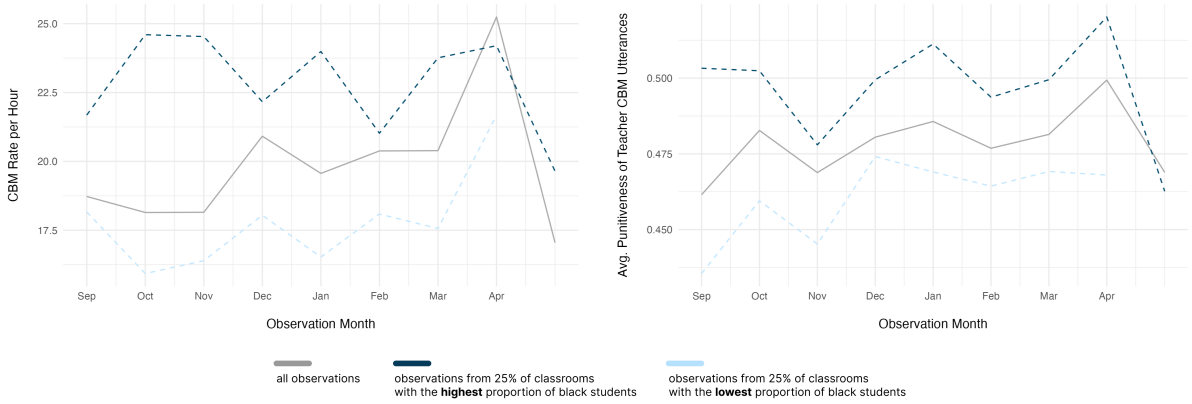

Figure 4: Variation in computed metrics over time across observation months in the school year.

classroom management language across observations throughout the academic year. Despite sharp fluctuates across months, both measures increase over the months of the school year. The punitiveness of classroom management increases steadily over the academic year. On average, classrooms observed at the end of the school year experience a 12.5% increase in the rate of classroom management language and demonstrate classroom management language

that is 4.7% more punitive than classrooms observed at the beginning of the year. These trends support prior findings that CLASS observation scores decrease over the school year (Casabianca et al., 2015). Classrooms with the highest proportion of Black students begin the school year with significantly more and more punitive classroom management language. They maintain a relatively constant trend, compared to more prominent increases in the general trend and that for classrooms with the lowest proportion of Black students.