Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 23-844

Do As I Say: What Teachers' Language Reveals About Classroom Management Practices

Mei Tan
Stanford University

Dorottya Demszky
Stanford University

Classroom management critically affects students' academic and behavioral outcomes, yet we lack quantitative methods for observing these practices at scale. This study develops and validates language-based measures of classroom management—such as responding to student behavior and issuing verbal or material sanctions—using natural language processing (NLP) on 1,652 elementary mathematics classroom transcripts. On average, classroom management language comprises 24% of teacher talk, and behavior management 7%, with wide variation across teachers. Novice teachers use more reprimands, while command frequency increases with experience. Classrooms with more Black students experience more language reflecting exclusionary discipline. These measures offer a new lens for studying and supporting classroom management as it unfolds in real time.

VERSION: November 2025

Suggested citation: Tan, Mei, and Dorottya Demszky. (2025). Do As I Say: What Teachers' Language Reveals About Classroom Management Practices. (EdWorkingPaper: 23-844). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/9yj6-jn52

Do As I Say:

What Teachers' Language Reveals About Classroom Management Practices

Mei Tan and Dorottya Demszky

Stanford University Graduate School of Education

Author Note

Mei Tan https://orcid.org/0009-0007-3871-899X

Dorottya Demszky https://orcid.org/0000-0002-6759-9367

The authors are grateful to Julie Cohen, Heather Hill, Alvin Pearman, Jim Malamut, and Sarah Novicoff for their thought partnership throughout the development of this study. We also thank Sarah Levine, Heather Hill, Jing Liu, Sanne Smith, Juergen Dieber, and our anonymous reviewers for their generous feedback on earlier drafts. This work would not have been possible without the hard work of our teacher annotators.

Correspondence concerning this article should be addressed to Mei Tan, Stanford University, 485 Lausen Mall, Stanford, CA 94305. Phone: (979) 739-1545; Email: mxtan@stanford.edu

Dorottya Demszky can be reached at Dorottya Demszky, Stanford University, 485 Lausen Mall, Stanford, CA 94305. Phone: (650) 242-5505; Email: ddemszky@stanford.edu

Abstract

Classroom management critically affects students' academic and behavioral outcomes, yet we lack quantitative methods for observing these practices at scale. This study develops and validates language-based measures of classroom management—such as responding to student behavior and issuing verbal or material sanctions—using natural language processing (NLP) on 1,652 elementary mathematics classroom transcripts. On average, classroom management language comprises 24% of teacher talk, and behavior management 7%, with wide variation across teachers. Novice teachers use more reprimands, while command frequency increases with experience. Classrooms with more Black students experience more language reflecting exclusionary discipline. These measures offer a new lens for studying and supporting classroom management as it unfolds in real time.

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

Do As I Say:

What Teachers' Language Reveals About Classroom Management Practices

Classroom management critically affects students' academic and behavioral outcomes (Korpershoek et al., 2016) as well as teachers' burnout and attrition (Aloe et al., 2014; Friedman, 2000). These outcomes are shaped not only by schoolwide policies or formal discipline systems, but by the moment-to-moment interactions through which teachers manage their classrooms—how they communicate expectations, redirect behavior, and establish order in real time. Despite growing efforts to improve teacher preparation and reduce exclusionary discipline (Osher et al., 2010), punitive and exclusionary strategies remain common in U.S. classrooms and continue to drive inequitable student experiences (Okonofua et al., 2016; Steyer et al., 2024). Yet the field lacks scalable methods for examining how such practices are taken up in everyday classroom practice. Addressing this gap is essential for developing nuanced understandings of how classroom management practices unfold across educational contexts and for informing efforts to improve them.

The Language of Classroom Management

Classroom management refers broadly to the actions teachers take to create and maintain environments conducive to learning (Brophy, 1983; Evertson & Weinstein, 2013). These may include nonverbal actions, such as organization of the physical classroom environment, established routines, gestures, eye contact, and the positioning of oneself strategically in the classroom (Woolfolk & Brooks, 1985). Yet in practice, many core aspects of classroom management are observable through language. Actions such as explaining rules, maintaining student attention, monitoring and providing accountability for task progress, facilitating group interactions, and responding to student behavior are accomplished through what teachers say and how they say it (Doyle, 2013; Woolfolk & Brooks, 1985).

Scholars have discussed classroom management—and the punitive aspects of its language—through a wide range of conceptual frameworks. Some frame classroom management as a reflection of teachers' underlying beliefs or orientations. For example, Milner et al. (2018) framed classroom management through restorative and punitive approaches, where the former emphasizes positive relationships and opportunities to make amends, and the latter centers accountability through punishment and establishing guilt. Similarly, Walker (2009) described the dimensions of nurture and control as key drivers of teachers' classroom management styles, shaping whether teachers are seen as authoritative or authoritarian. Others have drawn linguistic distinctions between controlling and autonomy-supportive (Reeve et al., 2014), teacher-centered and student-centered (Osher et al., 2010), and punitive and empathic approaches to classroom management (Okonofua et al., 2016). These frameworks treat tone and affect in teachers' language as meaningful signals to infer attitudes and emotional stances.

Other frameworks have focused on the form and content of teachers' classroom management language. Sieber (1976) identified over 30 types of talk moves in a study of elementary classrooms, including verbal reprimands, manipulation of privileges, threats, and isolation. Of these, short desists—brief and general unobtrusive reminders to get back on track—are particularly common (Doyle, 2013; Morine-Dershimer, 2013). Others have referenced threats, criticisms, and directives (Reeve et al., 2014), as well as language reflecting exclusionary discipline, such as calling guardians, isolating students within the classroom, and sending students out of the classroom (Smart & Igo, 2010; Steyer et al., 2024). These perspectives highlight the potential for identifying and quantifying punitive management practices through linguistic features in classroom discourse.

Computational Measures of Teacher Language

Natural language processing (NLP)—the use of computational methods to analyze and model language—offers a scalable, unobtrusive method for analyzing human behavior, cognition, and affect (Boyd & Schwartz, 2021; Munin et al., 2025; Pennebaker et al., 2015). In education, researchers have used NLP to quantify a range of constructs in text data, including teachers' instructional strategies and student engagement in tutoring dialogues (Chen et al., 2019; Wang et al., 2024), teachers' biases in written disciplinary referrals (Markowitz et al., 2023), and students' conceptual understanding of academic content in their written work (Somers et al., 2021). With the release of large-scale classroom transcript datasets (Demszky & Hill, 2023; Suresh et al., 2022), an emerging body of work now applies NLP to classroom discourse to study teaching and learning as they unfold in real time.

Drawing on classroom transcript data, researchers have developed NLP measures of teachers' questioning strategies (Alic et al., 2022; Kelly et al., 2018), uptake of student ideas (Demszky et al., 2021; Stone et al., 2019), dialogic talk moves (Suresh et al., 2022), and growth-mindset supportive language (Hunkins et al., 2022), among other practices. These measures have been used both descriptively—to study discourse patterns across teachers and contexts—and evaluatively, by linking teacher language to external indicators of instructional quality and student outcomes (Demszky & Hill, 2023; J. Liu & Cohen, 2021). They have also been used to support teacher professional learning directly as components of automated feedback tools (Demszky et al., 2023, 2025; Demszky & Liu, 2023; Jacobs et al., 2022). These NLP tools can thus be cost-effective complements to manual observation and enrich our understanding of instructional practice. Notably, however, no studies have yet applied these methods to analyze how teachers manage classrooms and student behavior.

The Present Study

Our study presents the first quantitative, large-scale analysis of classroom management language. We develop and validate a set of language-based measures of classroom management and apply them to 1,652 elementary mathematics classroom transcripts. We examine how teachers vary in their use of classroom management language, how that variation relates to teacher and classroom characteristics, and how it aligns with established observational and survey-based indicators of classroom climate. In doing so, we aim to both broaden the methodological toolkit available for studying classroom management and surface new insights into the everyday discourse through which discipline is enacted. To guide the work, we ask the following research questions:

1. How do teachers vary in their use of classroom management language?
2. How are differences in classroom management language associated with teacher identities, and classroom composition?
3. How do language-based measures of classroom management align with established observation- and survey-based indicators of classroom organization and climate?

Method

Our analytic workflow combined human annotation, NLP, and statistical analyses to quantify teachers' classroom management language. We developed a coding scheme to capture distinct forms of classroom management language—referred to here as *talk moves*—and applied it to annotate teacher utterances in a sample of transcripts. Based on the annotated examples, we trained and validated text classification models to identify these talk moves across the full dataset. We used model-predicted labels to construct frequency-based measures of language use for downstream analyses.

Data

Our analyses draw on a publicly released, multidimensional dataset collected by the National Center for Teacher Effectiveness (NCTE) between 2010 and 2013 (Kane et al., 2015). The dataset includes observations of 317 fourth- and fifth-grade teachers across 53 schools in New England that were primarily serving low-income students of color. Across classrooms, 67% of students were eligible for free or reduced-price lunch, 42% of students were Black and 24% were Hispanic. The teacher sample closely mirrors national patterns (National Center for Education Statistics, 2023), with 81% identifying as White and 83% as female.

Our dataset includes transcripts derived from 1,652 approximately 45-minute mathematics classroom observations (Demszky & Hill, 2023). During the original transcription process, names were anonymized using monikers such as “Student J,” “Teacher,” or “Mrs. H.” Transcribers segmented speech into utterances at their discretion, typically marking changes in speaker. To standardize units of analysis across observations, we further segmented each utterance at the sentence level and aligned it with video timestamps (see Appendix A for full preprocessing details). Our resulting analytic sample includes an average of 787 utterances per observation (*SD* = 262).

Associated with each transcript, the NCTE dataset includes observational ratings from the Classroom Assessment Scoring System (CLASS) protocol, a widely used observational tool that has demonstrated strong inter-rater reliability and predictive validity for assessing the quality of classroom interactions (Pianta, La Paro, & Hamre, 2008). CLASS evaluates three domains of teacher–student interaction: Classroom Organization (Behavior Management, Productivity, Negative Climate), Emotional Support (Positive Climate, Teacher Sensitivity, Regard for Student Perspectives), and Instructional Support (Instructional Learning Formats, Content Understanding, Analysis and Inquiry, Quality of Feedback, Instructional Dialogue, Student Engagement). Each dimension is rated on a 7-point scale by trained observers at 15-minute intervals of observation. In the NCTE dataset, instructional format—active instruction, small-group work, or a mix of both—was human-coded at 7.5-minute segments. The dataset also includes teacher and student demographic information, as well as survey measures of classroom climate (see Appendix B for additional metadata).

Human Annotation

Drawing on theoretical and observational work on classroom and behavior management (Doyle, 2013; Sieber, 1976; Steyer et al., 2024), we developed a hierarchical coding scheme of teachers' classroom management talk moves. At the highest level, the scheme distinguishes classroom management language from other forms of teacher talk. Following Doyle (2013), we define Classroom Management as the actions teachers take to create and sustain order. These actions range from organizing lessons to distributing resources, explaining rules, monitoring activities, and reacting to individual and group behavior. Within this domain, our coding scheme defines Behavior Management as the subset of Classroom Management language used to respond to student misbehavior. Although behavior management also includes proactive evidence-based strategies such as pre-corrections, opportunities to respond, and behavior-specific praise (Simonsen et al., 2008; Massar et al., 2023), our scheme specifically targets punitive strategies. Such Behavior Management is further categorized into verbal sanctions (Short Desists, Commands, Reprimands, Threats) and material sanctions, including Non-Exclusionary Consequences and Exclusionary Consequences (Calling Home, In-Class Isolation, Out-of-Class

Isolation). These talk moves were selected through iterative data review based on their conceptual relevance, reliance on observable linguistic features, and consistent identifiability in transcript data. Table 1 provides the full set of codes and definitions.

[Insert Table 1]

We recruited six experienced upper elementary mathematics teachers to annotate classroom transcripts at the utterance level. Annotators first practiced on a shared set of training examples, discussing disagreements as a group, and then proceeded to independently label examples to assess interrater reliability. Fleiss' Kappa values ranged from 0.719 to 0.964 for most talk moves (see Table 1), indicating substantial to near-perfect agreement. One category of talk moves—Nonexclusionary Consequences—showed moderate agreement (0.428) and was handled with more conservative annotation and modeling procedures (see Appendix C for details).

We conducted annotation in two phases. In the first phase, annotators labeled utterances for Classroom and Behavior Management (Table 1, Panel A). We selected 20 transcripts (10,355 utterances) from our data via stratified random sampling to reflect a range of CLASS scores on classroom organization dimensions (Behavior Management, Productivity, and Negative Climate). Each transcript was annotated in full to preserve conversational context. Because Behavior Management language is relatively infrequent, we used these annotations to train a preliminary classifier (using methods described in the next section) to identify likely Behavior Management utterances across the dataset and enable efficient sampling for the second phase.

In the second phase, annotators labeled utterances for the specific subset of verbal and material sanctions (Table 1, Panels B and C). We selected a random sample of 5,720 unique model-identified Behavior Management utterances, representing 200 observations. Annotators viewed each target utterance with three preceding and three following utterances for context. Additional details about sampling and annotation procedures are provided in Appendix C.

Model Training and Inference

We trained a series of text classification models to identify each talk move by fine-tuning RoBERTa (Y. Liu et al., 2019), a widely used transformer-based language model, on our annotated dataset. RoBERTa belongs to a class of large pretrained language models that come pre-equipped with context-specific representations of word meaning, acquired through their initial training on vast amounts of text data. These models can be fine-tuned on task-specific datasets to predict target constructs, reflecting the labeling judgments made by expert annotators. We fine-tuned separate RoBERTa models for each talk move using default parameters, with adjustments for particularly rare categories to ensure sufficient training examples for the model.

We evaluated each model using five-fold cross-validation—training on four partitions of the annotated dataset and testing on the fifth—with performance averaged across folds. F1 scores ranged from 0.862 to 0.974, indicating strong performance across talk moves. Final models were then trained on the full annotated dataset and used to generate predictions across the entire corpus of transcripts. Full model training and evaluation details are provided in Appendix D.

Analytic Approach

To address our first research question, we examined distributional patterns in the use of talk moves across observations and teachers. We calculated the percentage of total teacher utterances per observation dedicated to each talk move, as well as the percentage of teachers who used each talk move at least once. To examine sources of variation in language use, we also estimated linear mixed-effects models to decompose the variance in the frequency of each talk move across schools, teachers, and instructional format in 7.5-minute segments.

To address our second research question, we estimated linear regression models to examine the association between teacher and classroom characteristics (independent variables) and the hourly rate of each talk move (dependent variable). Models included controls for overall number of teacher utterances and total utterances per observation, as well as fixed effects for school, school year, and grade level. To assess their unique associations with classroom management language, all teacher and classroom demographic covariates were included in each model. Standard errors were clustered at the teacher level.

Finally, to address our third research question, we explored how our language measures relate to existing indicators of classroom and behavioral management. We computed partial Spearman correlations between teachers' hourly rates of each talk move and human-coded CLASS observational scores, as well as survey responses from teachers and students. To isolate these associations from contextual differences in classroom talk, we residualized all variables with respect to the number of teacher utterances and total utterances per observation, and included fixed effects for school, school year, and grade level.

Limitations

Our data and methods limit the inferences we can draw from this study. First, our models operate exclusively on text and do not capture the full multimodal and relational context of classroom interaction. Our analyses reflect only what is verbalized and transcribed, without access to tone, facial expression, gesture, proximity, or rapport between speakers. Second, our measures rely on human annotation, which necessarily involves subjective interpretation of speaker intent and situational dynamics. The talk moves themselves also encompass a wide range of affective tones and social functions. Given these limitations, we aim not to make normative claims about the quality or appropriateness of teachers' classroom management language, but simply to operationalize observable patterns to inform future investigation.

Finally, our sample represents a particular education context recorded between 2010 and 2013. Though research suggests that teaching practices have remained relatively stable over time (D. K. Cohen & Mehta, 2017; Cuban, 1993), shifts in policy and sociocultural context may influence how classroom management is enacted today. Likewise, the transcripts in our dataset are derived from fourth- and fifth-grade mathematics classrooms in 53 schools in the northeastern United States—our findings may not generalize to other subjects, grade levels, or geographic regions.

Results

We present results in three sections aligned with our research questions. First, we describe patterns in teachers' use of classroom management language. Second, we report associations between language use and teacher and classroom characteristics. Finally, we explore how these language-based measures align with observational and survey indicators of classroom and behavior management.

Descriptive Patterns of Classroom Management Language

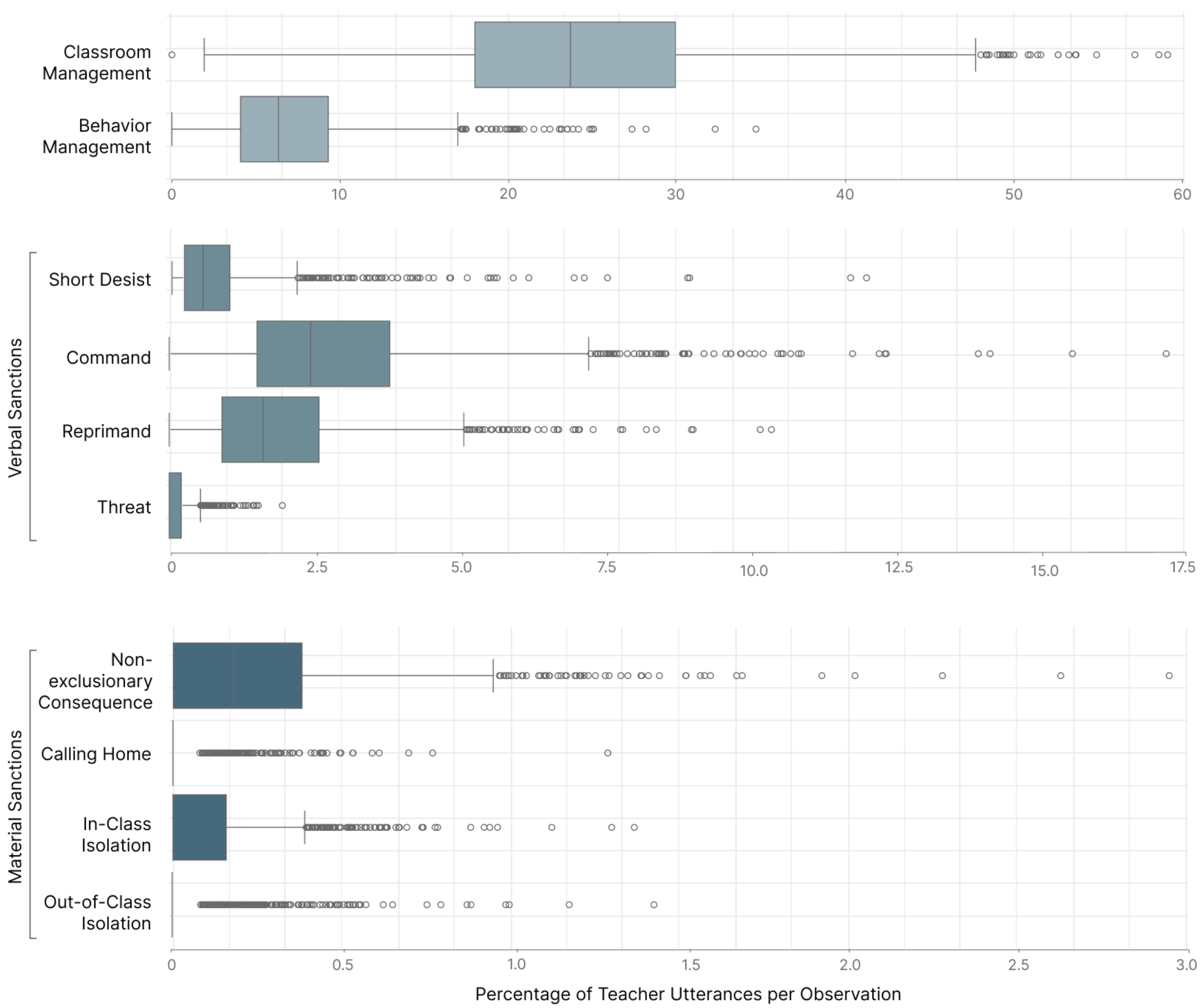
As shown in Figure 1, observations varied widely in how frequently Classroom and Behavior Management talk moves appeared in teacher talk. On average, Classroom Management accounted for 24% (*SD* = 9%) of all teacher utterances per observation, with proportions ranging from near zero to over 50%. Behavior Management accounted for 7% (*SD* = 4%) of teacher utterances on average, with some observations exceeding 30%.

Among types of verbal sanctions, Commands were most frequent (2.9%), followed by Reprimands (1.9%) and Short Desists (0.7%). Nonexclusionary Consequences accounted for

0.3% of utterances. More severe forms of disciplinary language, including Threats and Exclusionary Consequences—In-Class or Out-of-Class Isolation, and references to Calling Home—were rare on average (0.05%) but highly skewed. Some observations included rates exceeding 1% of teacher utterances, equivalent to roughly 8 or more instances in a single class period.

Figure 1

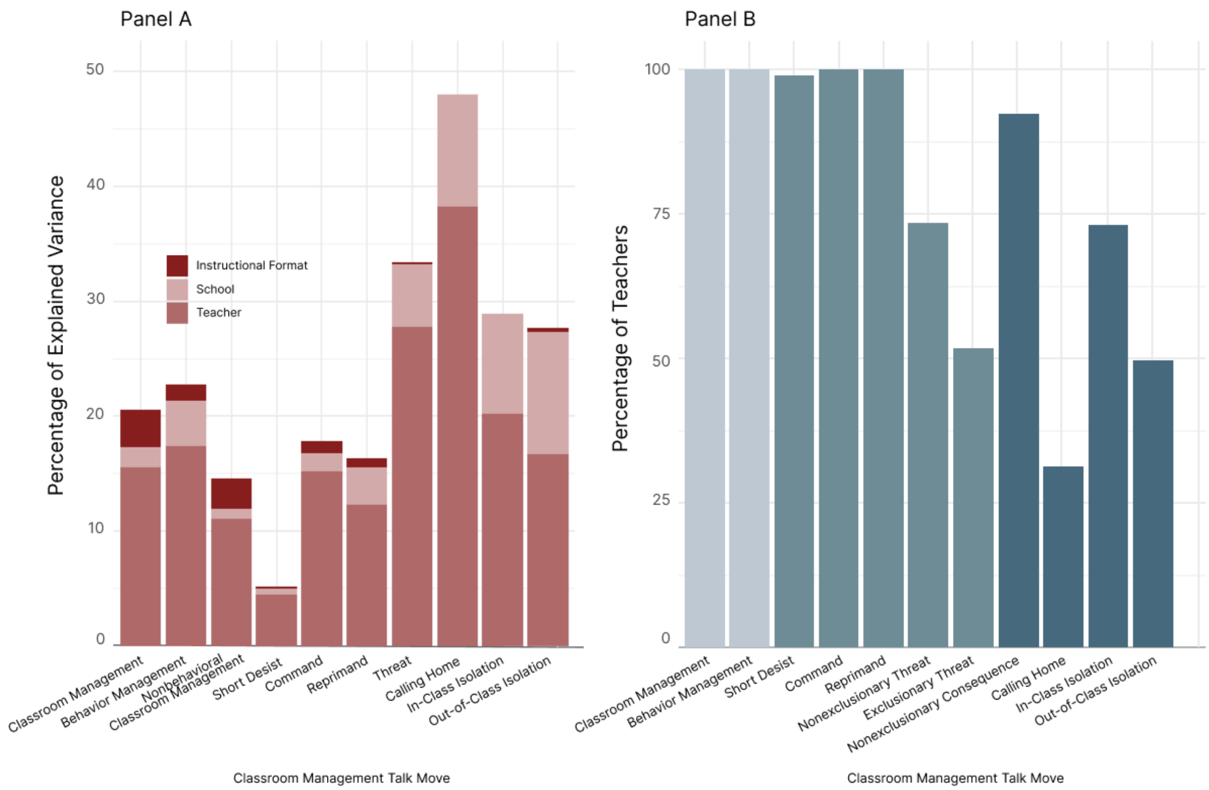
Percentage of Classroom Management Talk Moves Among Total Teacher Utterances per Observation



Across all talk moves, the largest share of explained variance occurred at the teacher level—16.4% on average across talk moves (Figure 2, Panel A). This suggests that teachers differ systematically in their use of classroom management language, even when accounting for shared school context (average 4.2%) and the instructional format of a segment (average 0.9%). However, sources of variation differed considerably across talk moves. Routine, nonexclusionary moves—such as the subset of Classroom Management unrelated to Behavior Management, Short Desists, and Nonexclusionary Consequences—had lower between-teacher variance (1.4% to 11.1%) and higher variance based on instructional format. The largest share of variance attributable to instructional format (2.7% to 3.4 %) was observed for general and nonbehavioral Classroom Management talk.

Figure 2

The Percentage of Explained Variance for Each Talk Move (Panel A) and The Percentage of Teachers Who Used Each Talk Move (Panel B)



Note. Nonbehavioral Classroom Management refers to the subset of Classroom Management that is not Behavior Management. Nonexclusionary Threats refers to those not reflecting Exclusionary Consequences.

In contrast, teacher-level differences accounted for 17.4% of the variance in overall Behavior Management language and 12.3% in Reprimands, with school-level contributions of 3.9% and 3.2%, respectively. Notably, the highest levels of between-teacher variation appeared in exclusionary talk moves: Threats (27.7%) and references to Calling Home (38.3%), In-Class Isolation (20.2%), and Out-of-Class Isolation (16.6%) showed the strongest teacher-level effects, along with the largest school-level contributions (5.5% to 10.6%). Full variance decomposition results are reported in Appendix E.

To further illustrate between-teacher variation, Panel B of Figure 2 shows the percentage of teachers observed using each talk move. Use of overall Classroom and Behavior Management language, Short Desists, Commands, Reprimands, and Nonexclusionary Consequences were nearly universal. In contrast, more severe forms of disciplinary language were used by a smaller subset of teachers. While 74% of teachers used threats of some kind, only 53% used threats that referenced exclusionary actions. Seventy-three percent of teachers used In-Class Isolation, 50% used Out-of-Class Isolation, and 31% referenced Calling Home.

Associations With Teacher and Classroom Characteristics

Differences in behavior management language were strongly associated with teachers' experience (see Table 2). Novice teachers (those with two years of experience or less) used, on average, eight more instances of Behavior Management language per hour than their more experienced peers—a 12% increase relative to the sample mean ($p < 0.01$). Novice teachers also issued 2.9 more Reprimands (+16%, $p < 0.001$), 2.5 more Commands (+9%, $p < 0.05$), and 1.3 more Short Desists (+18%, $p < 0.05$) per hour. Among non-novice teachers, each additional year of experience was associated with 0.4 more Commands (+1.3%, $p < 0.001$). While a positive coefficient also emerged for overall Behavior Management language (+0.7%, $p < 0.05$), visual inspection (Appendix F) suggests a non-linear trend, with higher rates among teachers with more than 25 years of experience.

[Insert Table 2]

Male teachers used less classroom management language across nearly all talk moves. They employed 20 fewer instances of Classroom Management language per hour (-9%, $p < 0.001$) and 9.5 fewer Behavior Management utterances (-14%, $p < 0.001$). They also issued 3.7 fewer Commands (-13.3%, $p < 0.001$), 2.9 fewer Reprimands (-15.9%, $p < 0.001$), and 20% fewer instances of language related to Out-of-Class Isolation (-0.17, $p < 0.001$) per hour relative to the sample mean. There was no notable difference for teachers who identified as Black. Small sample sizes limit our ability to analyze differences for other racial/ethnic groups and intersectional categories.

Holding teacher characteristics constant, larger class sizes were associated with more frequent Behavior Management language: Each additional student predicted 15.9 more instances of Behavior Management language per hour ($p < 0.001$), a 23% increase relative to the sample mean. Teachers in larger classes also issued 5.8 more Commands (+20.9%, $p < 0.001$) and 3.9 more Reprimands (+21.5%, $p < 0.001$) per hour. Interestingly, no significant associations were found between class size and the use of exclusionary talk moves.

Several student composition variables were also associated with differences in teacher language. Holding teacher characteristics constant, a 10-percentage-point increase in the proportion of male students was associated with 2 more Behavior Management utterances per hour (+2.9%, $p < 0.001$) and 0.8 more Commands (+2.9%, $p < 0.01$). A 10-percentage-point increase in the proportion of Black students was similarly associated with 1.9 more Behavior Management utterances (+2.8%, $p < 0.01$) and 1.2 more Commands (+4.3%, $p < 0.001$) per hour. Classrooms with a higher proportion of Black students also saw a modest but statistically significant increase in both exclusionary and nonexclusionary material sanctions. A 10-percentage-point increase in the proportion of Black students was associated with 7.6% more utterances related to In-Class Isolation relative to the sample mean (+0.07, $p < 0.001$).

Alignment With Observational and Survey Indicators

Finally, we examined how our language-based measures aligned with established indicators of classroom and behavior management. As illustrated in Figure 3, these measures showed strong, systematic correlations with CLASS observation scores. These theoretically consistent associations offer support for the construct validity of the measures while also pointing to their utility as complementary indicators that can add nuance to interpretation of broader observational frameworks.

Figure 3

Partial Spearman Correlations Between the Rate of Classroom Management Talk Moves per Hour, CLASS Observation Scores, and Survey Responses

Note. Correlations are computed on residuals from models that partial out the rates of teacher and total utterances per observation, as well as fixed effects for school, school year, and grade level. **p* < 0.05, ***p* < 0.01, ****p* < 0.001.

Teachers who used more Behavior Management language had lower CLASS scores across domains of Classroom Organization and Emotional Support, with the strongest correlations observed for Behavior Management (ρ = -0.29), Productivity (ρ = -0.24), and Negative Climate (ρ = 0.24). Reprimands were especially associated with elevated Negative Climate scores (ρ = 0.28) and lower Behavior Management scores (ρ = -0.26). Threats were similarly correlated with higher Negative Climate (ρ = 0.27), while Commands were associated with lower Behavior Management scores (ρ = -0.25). Correlations between talk moves and most CLASS scores were statistically significant ($p < 0.001$), including modest but reliable associations for In-Class and Out-of-Class Isolation (ρ = 0.09 to 0.12).

Language-based measures also aligned with teacher-reported experiences of classroom behavior. Teachers' reported frequency of reprimanding students correlated most strongly with their use of verbal sanctions, including Short Desists (ρ = 0.11), Reprimands (ρ = 0.10), and Threats (ρ = 0.11). Their reported loss of time due to misbehavior correlated most strongly with exclusionary language (ρ = 0.12), particularly In-Class Isolation (ρ = 0.10) and Calling Home (ρ = 0.07). Teachers feeling disrespected was most strongly correlated with Threats (ρ = 0.14), Reprimands (ρ = 0.09) and exclusionary language (ρ = 0.10), particularly In-Class Isolation (ρ = 0.09). Correlations with student-reported behavior were smaller in magnitude (ρ ≤ 0.09) but directionally consistent. Full regression results are reported in Appendix G.

Discussion

This study developed and validated language-based measures of classroom management using NLP methods applied to 1,652 classroom transcripts. Our descriptive and regression results show that teachers vary widely in their use of classroom and behavior management language, with between-teacher differences accounting for the largest share of explained variance across all talk moves. In line with prior research showing that behavior management is a major source of stress for new teachers (Aloe et al., 2014; Ingersoll & May, 2012), we found that novice teachers not only used significantly more Behavior Management language overall but also relied more heavily on reprimands. By contrast, the use of commands increased with each additional year of experience in the classroom.

These findings are perhaps consistent with research characterizing novice teachers as more likely to adopt reactive practices early in their careers (Reupert & Woodcock, 2010). However, contrary to research that found that novices are more likely to employ harsh interventions (Kwok, 2019; Tulley & Chiu, 1995), we observed no systematic changes in the use of exclusionary language with teacher experience. The persistence of exclusionary language across experience levels may signal the entrenchment of particular disciplinary approaches (Reupert & Woodcock, 2010; Walker, 2009), attitudes, and beliefs (Hoy & Weinstein, 2013; Kwok, 2017), or suggest that even experienced teachers may benefit from professional learning focused on more constructive, relational approaches to discipline (Milner & Laughter, 2015; Stough, 2013).

Even after controlling for teacher characteristics, classrooms with higher percentages of Black and Hispanic students received more frequent references to in-class and out-of-class isolation. While our data do not allow us to identify the recipients of these utterances, the presence of these associations raises important concerns. These findings are consistent with a wide body of research documenting racial disparities in classroom discipline (Gregory et al., 2010; Skiba et al., 2011). They suggest that such disparities extend to the everyday language of behavior management, not just formal referrals, echoing recent work on “de facto suspensions” that occur in classroom practice (Steyer et al., 2024).

Though modest in absolute terms, correlations with CLASS and teacher-reported survey items were statistically robust and directionally consistent, supporting the construct validity of the measures. The strength of these correlations is consistent with prior research linking observational features of classroom practice to student outcomes and other indicators of instructional quality (Pianta, Belsky, et al., 2008), and reflects meaningful variation in complex instructional settings (Kraft, 2020). Importantly, these measures offer a way to unpack the broad categories used in classroom observation frameworks by revealing the specific discourse practices that may inform—or be obscured within—evaluator judgments of instructional quality. Our findings suggest that such observational ratings are more strongly correlated with the overall frequency of classroom and behavior management talk than with the frequency of language reflecting exclusionary discipline. By distinguishing between types of disciplinary language, our language-based measures begin to tease apart the effects of frequency from severity, as frequent behavioral redirections do not necessarily indicate poor teaching or inefficient classroom management (Anagnostopoulos et al., 2020).

Implications for Practice

Our findings underscore the need for systems that attend not only to formal disciplinary outcomes, but also to the everyday language practices through which discipline is enacted—and through which teachers can learn. Tools built on our measures could support teachers in identifying moments of classroom and behavior management and reflecting on their practice. In classrooms, phone and web applications like TeachFX (https://teachfx.com/), TalkMoves (Suresh et al., 2021), and M-Powering Teachers (https://mpoweringteachers.stanford.edu/) already enable teachers to capture audio and receive automated feedback on their language. In teacher preparation and professional learning contexts, these measures could be integrated into technology-supported simulation environments (J. Cohen et al., 2020) or coaching platforms designed to strengthen teachers' real-time decision-making, such as EdThena (https://www.edthena.com/ai-coach-for-teachers/).

At the school and policy levels, language-based measures offer a scalable means of evaluating efforts to improve school discipline and reduce exclusionary practices. Disciplinary reforms are typically assessed with school-level outcomes rather than fine-grained signals at the level of individual classroom practice (Leung-Gagné et al., 2022; Lozen & Martinez, 2020; Osher et al., 2010). And while national rates of suspensions and expulsions have declined in recent years, exclusionary dynamics often persist in more subtle classroom interactions—patterns that language-based measures can make visible (Steyer et al., 2024; Gleit, 2023).

Importantly, these measures are feasible to implement. Audio can be recorded alongside existing observation protocols, and small cost-effective models can generate useful language-based analyses at scale. These analyses can complement human evaluator judgments by surfacing patterns of talk that are difficult to track in real time but central to how teachers manage classrooms.

Future Directions

This study opens several avenues for future research. One direction involves developing richer models of classroom management. Future work could capture variation in affect and tone within each measured talk move (Horsch et al., 2002). For example, commands and reprimands vary in the degree to which they sound harsh, shame, or fail to preserve student dignity (Sullivan et al., 2014). These nuances could be explored through the analysis of teacher affect in text as well as in voice and gestures in multimodal data (Wani et al., 2021). Additionally, transcription conventions that capture the directionality (e.g., the addressee) of each utterance would enable relational analyses of how classroom management language is distributed across students. While this study focused on a subset of classroom management talk moves related to punitive behavior management, future research could also model instructional routines and preventive strategies that support classroom order and student engagement.

A second direction for future research involves using language-based measures to inform the design of interventions. Applied across a wider range of classrooms, these measures could help explain how patterns of classroom management language vary with broader school conditions—such as school and disciplinary climate, policy environment, and student backgrounds—as well as with features of classroom instruction, including content area, activity structure, classroom layout, and instructional routines. Temporal analyses could model how patterns of management language unfold over the course of a lesson or evolve across the school year. These measures could also be used to examine how teachers' use of classroom management language relates to their preparation and professional learning experiences. By applying natural language processing to classroom transcripts, this work broadens the methodological toolkit for analyzing and supporting teachers' classroom management practice.

References

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 224–233). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bea-1.27

Aloe, A. M., Amo, L. C., & Shanahan, M. E. (2014). Classroom management self-efficacy and burnout: A multivariate meta-analysis. *Educational Psychology Review*, *26*(1), 101–126. https://doi.org/10.1007/s10648-013-9244-0

Anagnostopoulos, D., Cavanna, J., & Charles-Harris, S. (2020). Managing to teach ambitiously in the first year? *The Elementary School Journal, 120*(4), 667–691. https://doi.org/10.1086/708660

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Brophy, J. E. (1983). Classroom organization and management. *The Elementary School Journal*, *83*(4), 265–285. https://doi.org/10.1086/461318

Chen, G., Ferreira, R., Lang, D., & Gasevic, D. (2019). Predictors of student satisfaction: A large-scale study of human-human online tutorial dialogues. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 19–28). https://files.eric.ed.gov/fulltext/ED599229.pdf

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. https://doi.org/10.3102/0002831217700078

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis, 42*(2), 54–76. https://doi.org/10.3102/0162373720906217

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

Demszky, D., & Hill, H. (2023). The NCTE transcripts: A dataset of elementary math classroom transcripts. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 528–538). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.44

Demszky, D., & Liu, J. (2023). *M-Powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes* (EdWorkingPaper 23-759). Annenberg Institute at Brown University. https://doi.org/10.26300/s8xh-zp45

Demszky, D., Liu, J., Hill, H., Jurafsky, D., & Piech, C. (2023). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 224–233. https://doi.org/10.3102/01623737231169270

Demszky, D., Liu, J., Hill, H. C., Sanghi, S., & Chung, A. (2025). Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education*, *227*(1), 105183.

https://doi.org/10.1016/j.compedu.2024.105183

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. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1638–1653). https://doi.org/10.18653/v1/2021.acl-long.130

Doyle, W. (2013). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 107–136). Routledge. https://doi.org/10.4324/9780203874783

Evertson, C. M., & Weinstein, C. S. (2013). Classroom management as a field of inquiry. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 13–26). Routledge.

Gleit, R. D. (2023). Brokers and Boundary Managers: School Expulsions amid the Non-PunitiveTurn. Social Problems, spad025. https://doi.org/10.1093/socpro/spad025

Leung-Gagné, M., McCombs, J., Scott, C., & Losen, D. (2022). Pushed out: Trends anddisparities in out-of-school suspension. Learning Policy Institute. https://doi.org/10.54300/235.277

Losen, D. J., & Martinez, P. (2020). Is California Doing Enough to Close the School DisciplineGap? Center for Civil Rights Remedies at Civil Rights Project/Proyecto DerechosCiviles. https://www.civilrightsproject.ucla.edu/research/k-12-education/school-discipline/is-california-doing-enough-to-close-the-school-discipline-gap

Friedman, I. A. (2000). Burnout in teachers: Shattered dreams of impeccable professional performance. *Journal of Clinical Psychology*, *56*(5), 595–606. https://doi.org/10.1002/(SICI)1097-4679(200005)56:5%3C595::AID-JCLP2%3E3.0.CO;2-Q

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. https://doi.org/10.3102/0013189X09357621

Horsch, P., Chen, J.-Q., & Wagner, S. L. (2002). The responsive classroom approach: A caring, respectful school environment as a context for development. *Education and Urban Society*, *34*(3), 365–383. https://doi.org/10.1177/0013124502034003006

Hoy, A. W., & Weinstein, C. S. (2013). Student and teacher perspectives on classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 191–230). Routledge.

Hunkins, N., Kelly, S., & D'Mello, S. (2022, March 21–25). *"Beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms* [Conference session]. 12th International Learning Analytics and Knowledge Conference, New York, NY, United States. https://doi.org/10.1145/3506860.3506896

Ingersoll, R. M., & May, H. (2012). The magnitude, destinations, and determinants of mathematics and science teacher turnover. *Educational Evaluation and Policy Analysis*, *34*(4), 435–464. https://doi.org/10.3102/0162373712454326

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*, Article 103631. https://doi.org/10.1016/j.tate.2022.103631

Kane, T., Hill, H., & Staiger, D. (2015). *National Center for Teacher Effectiveness Main Study*. Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR36095.v4

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. https://doi.org/10.3102/0013189X18785613

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. https://doi.org/10.3102/0034654315626799

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kwok, A. (2019). Classroom management actions of beginning urban teachers. *Urban Education, 54*(3), 339-367. https://doi.org/10.1177/0042085918795017

Kwok, A. (2017). Relationships between instructional quality and classroom management for beginning urban teachers. *Educational Researcher, 46*(7), 355-365. https://doi.org/10.3102/0013189X17726727

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. https://doi.org/10.3102/01623737211009267

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., &

Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. https://doi.org/10.48550/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*, *52*(9), 525–534. https://doi.org/10.3102/0013189X231189444

Massar, M. M., Horner, R. H., Kittelman, A., & Conley, K. M. (2023). Mechanisms of effective coaching: Using prompting and performance feedback to improve teacher and student outcomes. *Journal of Positive Behavior Intervention*s, 25(3), 169-184. https://doi.org/10.1177/10983007221133524

Milner, H. R., IV, Cunningham, H. B., Delale-O'Connor, L., & Kestenberg, E. G. (2018). *"These kids are out of control": Why we must reimagine "classroom management" for equity*. Corwin Press.

Milner, H. R., & Laughter, J. C. (2015). But good intentions are not enough: Preparing teachers to center race and poverty. *The Urban Review*, *47*(2), 341–363. https://doi.org/10.1007/s11256-014-0295-4

Morine-Dershimer, G. (2013). Classroom management and classroom discourse. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management* (pp. 137–166). Routledge.

Munin, S., Ong, D. C., Okland, S., Freedman, G., & Beer, J. S. (2025). Language measures correlate with other measures used to study emotion. *Communications Psychology*, *3*(1), Article 29. https://doi.org/10.1038/s44271-025-00212-x

National Center for Education Statistics. (2023). *Characteristics of public school teachers*. U.S.

Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/pdf/2023/clr_508.pdf

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 of the United States of America*, *113*(19), 5221–5226. https://doi.org/10.1073/pnas.1523698113

Osher, D., Bear, G. G., Sprague, J. R., & Doyle, W. (2010). How can we improve school discipline? *Educational Researcher*, *39*(1), 48–58. https://doi.org/10.3102/0013189X09357618

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. https://doi.org/10.13140/RG.2.2.23890.43205

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397. https://doi.org/10.3102/0002831207308230

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

Reeve, J., Vansteenkiste, M., Assor, A., Ahmad, I., Cheon, S. H., Jang, H., Kaplan, H., Moss, J. D., Olaussen, B. S., & Wang, C. K. J. (2014). The beliefs that underlie autonomy-supportive and controlling teaching: A multinational investigation. *Motivation and Emotion*, *38*(1), 93–110. https://doi.org/10.1007/s11031-013-9367-0

Reupert, A., & Woodcock, S. (2010). Success and near misses: Pre-service teachers' use, confidence and success in various classroom management strategies. *Teaching and*

Teacher Education, *26*(6), 1261–1268. https://doi.org/10.1016/j.tate.2010.03.003

Sieber, R. T. (1976). *Schooling in the bureaucratic classroom: Socialization and social reproduction in Chestnut Heights* (Publication No. 302799120) [Doctoral Dissertation, New York University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/schooling-bureaucratic-classroom-socialization/docview/302799120/se-2

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*, 31(3), 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. https://doi.org/10.1080/02796015.2011.12087730

Smart, J. B., & Igo, L. B. (2010). A grounded theory of behavior management strategy selection, implementation, and perceived effectiveness reported by first-year elementary teachers. *The Elementary School Journal*, *110*(4), 567–584. https://doi.org/10.1086/651196

Somers, R., Cunningham-Nelson, S., & Boles, W. (2021). Applying natural language processing to automatically assess student conceptual understanding from textual responses. *Australasian Journal of Educational Technology*, *37*(5), 98–115. https://doi.org/10.14742/ajet.7121

Steyer, L., Provencal, M., Pearman, F., & Obradović, J. (2024). *De facto suspensions: Informal early childhood exclusionary discipline practices in an era of formal school discipline reform*. EdArXiv. https://doi.org/10.35542/osf.io/vt7rf

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. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 420–425). https://eric.ed.gov/?id=ED599238

Stough, L. (2013). The place of classroom management and standards in teacher education. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management* (pp. 919–934). Routledge.

Sullivan, A. M., Johnson, B., Owens, L., & Conway, R. (2014). Punish them or engage them? Teachers' views of unproductive student behaviours in the classroom. *Australian Journal of Teacher Education*, *39*(6), Article 4. https://doi.org/10.14221/ajte.2014v39n6.6

Suresh, A., Jacobs, J., Clevenger, C., Lai, V., Tan, C., Martin, J. H., & Sumner, T. (2021). Using AI to promote equitable classroom discussions: The TalkMoves application. In I. Roll, D. McNamara, S. Sosnovsky, R. Jickin, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 344–348). Springer. https://doi.org/10.1007/978-3-030-78270-2_61

Suresh, A., Jacobs, J., Clevenger, C., Perkoff, M., Martin, J., & Sumner, T. (2022). The TalkMoves dataset: K–12 mathematics lesson transcripts annotated for teacher and student discursive moves. In N. Calzolari, R. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 13th Language Resources and Evaluation Conference* (pp. 4654–4662). European Language Resources Association. https://aclanthology.org/2022.lrec-1.497/

Tulley, M., & Chiu, L. H. (1995). Student teachers and classroom discipline. *The Journal of*

Educational Research, *88*(3), 164–171. https://doi.org/10.1080/00220671.1995.9941295

Walker, J. M. T. (2009). Authoritative classroom management: How control and nurturance work together. *Theory Into Practice*, *48*(2), 122–129. https://doi.org/10.1080/00405840902776392

Wang, R., Zhang, Q., Robinson, C., Loeb, S., & Demszky, D. (2024). Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2174–2199). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.120

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, *9*, 47795–47814. https://doi.org/10.1109/ACCESS.2021.3068045

Woolfolk, A. E., & Brooks, D. M. (1985). The influence of teachers' nonverbal behaviors on students' perceptions and performance. *The Elementary School Journal*, *85*(4), 513–528. https://doi.org/10.1086/461418

Table 1

Taxonomy of Classroom Management Talk Moves, With Definitions, Examples, and Interrater Agreement

Dimension	Definition	Examples	IRR
Panel A (Total Annotated Examples = 10,355)			
Classroom Management (Doyle, 2013)	Language that reflects teacher actions aimed at creating and maintaining an orderly environment conducive to academic learning. This includes organizing lessons, managing transitions, explaining rules, monitoring progress, facilitating group interactions, maintaining accountability, and redirecting attention or behavior to support instructional goals.	"Please stand up, push in your seats, and line up quietly"; "I want you to get your books; you may lay anywhere on the floor and read quietly"; "Make sure that you have your homework out to give to me as you're going out the door over there"; "You need to be in your seats by the count of five"	0.782
Behavior Management (Doyle, 2013)	A subset of Classroom Management language focused on responding to individual or group behaviors that initiate a competing vector of action—behaviors that threaten or distract from the instructional trajectory.	"I need voices off, and I need your eyes on the board"; "I'm only gonna call on people who are sitting down"; "Everyone needs to be listening"; "Student E, don't pack up"; "Stop that right now"	0.916
Panel B (Total Annotated Examples = 5,720)			
Verbal Sanction	A subset of Behavior Management that involves "only telling." These are verbal responses, statements or directives with no mention of material consequences (Sieber, 1976).		---
↳ **Short Desist** (Doyle, 2013; Sieber, 1976)	A subset of verbal sanctioning that involves brief, often unobtrusive verbal cues or reminders used to quickly stop misbehavior or redirect student attention back to task.	"Shh"; "Excuse me"; "Hold on"; "Focus"	0.719
↳ **Command** (Sieber, 1976)	A subset of verbal sanctioning that involves a direct command issued to an individual student to change behavior, typically without explicitly expressing disapproval.	"I need you to have a seat"; "Put your hand down"; "Student D, put that away"; "I don't want to hear any talking"	0.723
↳ **Reprimand** (Sieber, 1976)	A subset of verbal sanctioning that conveys teacher disapproval and expressions of norm violation, often with a reproving tone; may involve mocking or ridicule.	"Student H, you're being rude"; "You still didn't do the very first thing, Student K. Student F, is your name Student K?"; "I'm tired of your laziness"; "Student A, there's no reason to get that loud. Stop yelling out"	0.785
↳ **Threat** (Sieber, 1976)	A subset of verbal sanctioning that implies or states the possibility of a future material consequence.	"Don't make me ask again"; "Student T and Student E, this is my warning before I separate you"; "If your body is not turned towards me, I'm assuming you're not paying attention and you will get a	0.901

		consequence"; "Student A, you're gonna leave if you do that again"	
Panel C (Total Annotated Examples = 5,720)			
Material Sanction	A subset of Behavior Management involving consequences that are "more than telling." These include manipulations of access to material goods or changes to bodily or social states (Sieber, 1976).		---
↳ **Non-exclusionary Consequences** (Sieber, 1976)	A subset of material sanctions in which student participation in classroom activities is preserved. This includes manipulation of token systems, removal of privileges or items, and other disciplinary actions that do not involve social or spatial isolation.	"Student G, give me those"; "I'll take that. Do not have this in the class again"; "Please move your clothespin"; "This shows me that you're not paying attention when I'm up here teaching so you need to bring me your passport now, please"; "You have a worksheet now because clearly we're not doing what we're supposed to be doing at the promethium board"	0.428
↳ **Exclusionary Consequences**	A subset of material sanctions that involves Calling Home, In-Class Isolation, or Out-of-Class Isolation.		---
↳ **Calling Home** (Sieber, 1976; Steyer, 2024)	A subset of material sanctions that involve contacting a student's parent or guardian to report misbehavior.	"Your mother will not be happy"; "I can call home, too, as well"; "I talked to your father this morning, and there you are, acting like a fool"; "You know what, Student A, I don't want to have to show your parents that you're fooling around"	0.964
↳ **In-Class Isolation** (Sieber, 1976; Steyer, 2024)	A subset of material sanctions in which a student is spatially or socially separated from the classroom group while remaining in the classroom setting. This includes moving the student to a separate area or withholding participation in group activities.	"Student D, do you need to not participate?"; "Can you actually take all of yours and I'm gonna have you sit in the back of the room please"; "Okay Student M and E, if I see that again you will be sitting out"; "Go to time out take your stuff and go over there—I can't take that noise."	0.831
↳ **Out-of-Class Isolation** (Sieber, 1976; Steyer, 2024)	A subset of material sanctions that involve removing the student from the classroom space entirely, such as sending them to the hallway or to another room with a noninstructional adult.	"Student D, and Student E, please go outside my door"; "Student C, you can go out in the hall"; "If this continues, you're going to be asked to leave"; "Student L, go physically stand outside of the doorway"	0.940

Note. Interrater reliability (IRR) is measured with Fleiss' Kappa. Missing IRR values indicate higher-level dimensions for which only subdimensions were coded.

Table 2

Correlations Between the Hourly Rate of Classroom Management Talk Moves and Teacher and Student Characteristics

Predictor	(1) Classroom Management	(2) Behavior Management	(3) Short Desist	(4) Command	(5) Reprimand	(6) Threat	(7) Non-exclusionary Consequence	(8) Exclusionary Consequence	(9) Calling Home	(10) In-Class Isolation	(11) Out-of-Class Isolation
Teacher Characteristics											
Years of experience	0.537	0.505***	0.070*	0.353***	0.009	-0.012*	-0.016*	-0.003	0.002	-0.003	-0.000
	(0.362)	(0.176)	(0.04)	(0.085)	(0.047)	(0.006)	(0.008)	(0.007)	(0.002)	(0.005)	(0.003)
Novice	8.494	8.039**	1.293*	2.537*	2.888***	0.315	0.371	0.268	0.122*	0.163	0.042
	(6.885)	(3.125)	(0.78)	(1.474)	(1.021)	(0.207)	(0.254)	(0.169)	(0.065)	(0.121)	(0.066)
Male	-19.956***	-9.453***	-0.266	-3.746***	-2.873***	-0.130	-0.441***	-0.244**	-0.014	-0.173***	-0.051
	(5.686)	(3.036)	(0.724)	(1.428)	(0.911)	(0.124)	(0.146)	(0.106)	(0.043)	(0.064)	(0.051)
Black	10.312	3.150	0.138	1.728	0.545	-0.015	-0.224	0.092	0.102*	-0.058	0.111*
	(7.074)	(3.486)	(0.48)	(1.797)	(1.097)	(0.126)	(0.138)	(0.139)	(0.054)	(0.071)	(0.061)
Class Characteristics											
Class size	23.835*	15.857***	3.083**	5.816**	3.935**	0.039	0.165	0.059	0.106	-0.061	-0.004
	(12.531)	(5.825)	(1.409)	(2.640)	(1.540)	(0.189)	(0.381)	(0.188)	(0.071)	(0.131)	(0.094)
% Male students	3.395**	1.965***	0.180	0.778**	0.439*	0.020	0.038	0.044	-0.002	0.020	0.025
	(1.561)	(0.745)	(0.169)	(0.340)	(0.227)	(0.040)	(0.045)	(0.035)	(0.011)	(0.021)	(0.018)
% Black students	2.561	1.891**	0.248	1.196***	0.202	0.051	0.124**	0.084***	-0.002	0.066***	0.016
	(1.671)	(0.817)	(0.174)	(0.369)	(0.234)	(0.033)	(0.050)	(0.032)	(0.012)	(0.023)	(0.015)
% Hispanic students	1.104	1.003	-0.016	1.107***	0.033	0.014	0.053	0.100***	-0.008	0.076**	0.034**
	(2.338)	(0.904)	(0.216)	(0.422)	(0.270)	(0.034)	(0.056)	(0.039)	(0.016)	(0.031)	(0.017)
% FRPL students	-2.329	-0.290	0.075	-0.180	-0.075	-0.010	-0.033	-0.059*	-0.021*	-0.018	-0.032**
	(1.600)	(0.792)	(0.217)	(0.371)	(0.210)	(0.031)	(0.049)	(0.034)	(0.012)	(0.024)	(0.016)
% ELL students	0.079	-0.048	-0.017	-0.108	-0.092	-0.007	-0.007	-0.038*	-0.006	-0.041***	0.000
	(1.172)	(0.692)	(0.115)	(0.301)	(0.211)	(0.022)	(0.037)	(0.020)	(0.008)	(0.014)	(0.009)
% SPED students	1.604	0.008	-0.064	0.181	-0.242	-0.006	0.087*	-0.014	0.005	-0.027	0.010
	(1.64)	(0.823)	(0.17)	(0.386)	(0.256)	(0.030)	(0.044)	(0.036)	(0.014)	(0.02)	(0.014)
Mean	235	69.214	7.325	27.837	18.107	1.33	2.437	1.302	0.195	0.869	0.377

Note. $N = 1543$. FRPL = free or reduced-price lunch; ELL = English-language learners; SPED = special education. Columns display results from separate regressions. Each model includes school, school year, and grade-level fixed effects, as well as controls for the total number of utterances and teacher utterances per observation. Class size is log-transformed. Coefficients for student demographic variables reflect the effect of a 10-percentage-point increase. Robust standard errors are clustered at the teacher level.

$^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Figure Captions

Figure 1

Percentage of Classroom Management Talk Moves Among Total Teacher Utterances per Observation

Figure 2

The Percentage of Explained Variance for Each Talk Move (Panel A) and The Percentage of Teachers Who Used Each Talk Move (Panel B)

Note. Nonbehavioral Classroom Management refers to the subset of Classroom Management that is not Behavior Management. Nonexclusionary Threats refers to those not reflecting Exclusionary Consequences.

Figure 3

Partial Spearman Correlations Between the Rate of Classroom Management Talk Moves per Hour, CLASS Observation Scores, and Survey Responses

Note. Correlations are computed on residuals from models that partial out the rates of teacher and total utterances per observation, as well as fixed effects for school, school year, and grade level. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Appendix A

Data Preprocessing Procedures

The publicly released NCTE transcripts segment speech according to transcriber discretion, often aligning with speaker changes but not consistently delimiting individual sentences. For this study, we obtained a post-processed version of the transcripts with sentence-level timestamps to enable more consistent linguistic analysis. Utterances had been segmented at the sentence level using the Natural Language Toolkit (NLTK) sentence tokenizer. These sentence-level utterances were force-aligned with the original classroom video data, enabling accurate timestamping at the sentence level. This process produced aligned data for 1,652 of the original 1,660 classroom observations. The resulting dataset—which we used in our analyses— included 1,843,623 sentence-level utterances (*M* = 787 per observation; *SD* = 262).

We also identified utterances consisting of short discourse markers (e.g., "okay," "alright," "so," "well") that function as pragmatic cues but carry limited semantic content on their own. These utterances were identified using a predefined list of common discourse markers and were excluded from model training, inference, and downstream analyses. They were not removed from the transcripts, and annotators viewed them as part of the full transcript context during labeling.

Appendix B

Variables From the NCTE Dataset Used in Study Analyses,

Including Percentage of Missing Values, Mean, and Standard Deviation Values

Variable	Description	% missing	*M*	*SD*
Teacher Characteristics				
Years of experience	Teacher's years of experience, including school year of survey.	0.80	10.55	6.86
Male	Binary indicator for whether the teacher is male.	0.00	0.17	0.38
Black	Binary indicator for whether the teacher is Black.	0.00	0.19	0.39
Hispanic	Binary indicator for whether the teacher is Hispanic.	0.00	0.03	0.18
Classroom Characteristics				
Class size	Number of students in the classroom at the time of observation.	1.80	20.9	5.08
% Male students	Binary indicator for whether a student is male, aggregated by class identifier at the observation level.	1.80	0.17	0.13
% Black students	Binary indicator for whether a student is Black, aggregated by class identifier at the observation level.	1.80	0.42	0.26
% Hispanic students	Binary indicator for whether a student is Hispanic, aggregated by class identifier at the observation level.	1.80	0.24	0.23
% FRPL students	Binary indicator for whether a student received free or reduced-price lunch in the year of observation, aggregated by class identifier at the observation level.	1.80	0.67	0.26
% SPED students	Binary indicator for whether a student hads special education status in the year of observation, aggregated by class identifier at the observation level.	1.80	0.14	0.17
% LEP students	Binary indicator for whether a student had limited English proficiency in the year of observation, aggregated by class identifier at the observation level.	1.80	0.22	0.25
CLASS Observation Measures				
Behavior management	Classroom organization dimensions of the CLASS observational instrument. Scores range from 1 (*low*) to 7 (*high*). Scored in 15-minute segments and mean aggregated at the observation level.	0.20	6.09	0.84
Productivity		0.20	6.36	0.73
Negative climate		0.20	5.25	0.96
Positive Climate	Emotional support dimensions of the CLASS observational instrument. Scores range from 1 (*low*) to 7 (*high*). Scored in 15-minute segments and mean aggregated at the observation level.	0.20	1.18	0.41
Teacher sensitivity		0.20	4.61	0.82
Regard for student perspectives		0.20	3.53	0.96
Teacher Survey Measures				
Frequency of reprimanding students	Teacher survey items. Values range from 1 (*rarely or never*) to 5 (*always*).	2.10	2.13	0.97
Frequency of losing time to student misbehavior		2.10	1.86	0.99

Frequency of feeling disrespected		2.10	1.46	0.8
Student Survey Measures				
My behavior in this class is good	Student survey items for student agreement with the sentiment. Values range from 1 (*totally untrue*) to 5 (*totally true*).	7.14	4.23	0.89
My behavior in this class sometimes annoys the teacher		8.84	2.23	1.36
My behavior is a problem for the teacher in this class		8.59	1.75	1.15
Instructional Format				
Format: Active instruction	Scored 1 when teacher leads discussion, presentation of mathematical material, or posing of mathematical problems to the group.	21.03	0.46	0.46
Format: Small-group work	Scored 1 when teacher divides students into small groups or pairs, or students work individually on a mathematical problem or task.	21.03	0.20	0.36
Format: Both	Scored 1 when at least one minute of each type of instruction (active instruction and small-group work) occurs in the segment.	21.03	0.33	0.40

Note. FRPL = free or reduced-price lunch; ELL = limited English proficiency; SPED = special education.

Appendix C:

Annotation Procedures

We selected six experienced upper elementary mathematics teachers as annotators. Three had over 8 years of classroom experience, and three had at least 3 years. The annotators included five women and one man, and identified as Asian (2), Hispanic (1), and White (3). All were affiliated with our university's school of education—as alumni of the teacher preparation program, participants in a professional development fellowship, or members of a research-practice network supporting teacher engagement in research.

For all annotation tasks, teachers were provided with a detailed codebook that included definitions, examples, and nonexamples for each coded dimension. In Phase 1, we selected 20 transcripts for annotation using stratified random sampling to ensure a range of CLASS scores. For the Behavior Management dimension, we sampled four low-scoring observations (< 3) and four mid-to-high scoring observations (≥ 3); for Negative Climate and Productivity, three low and three mid-to-high scoring observations were selected. Annotators jointly reviewed one transcript as a training set, resolving disagreements through group discussion and consensus. Teachers then independently annotated two transcripts to establish interrater reliability. Annotators then received individualized calibration feedback and proceeded to annotate three to four additional transcripts independently.

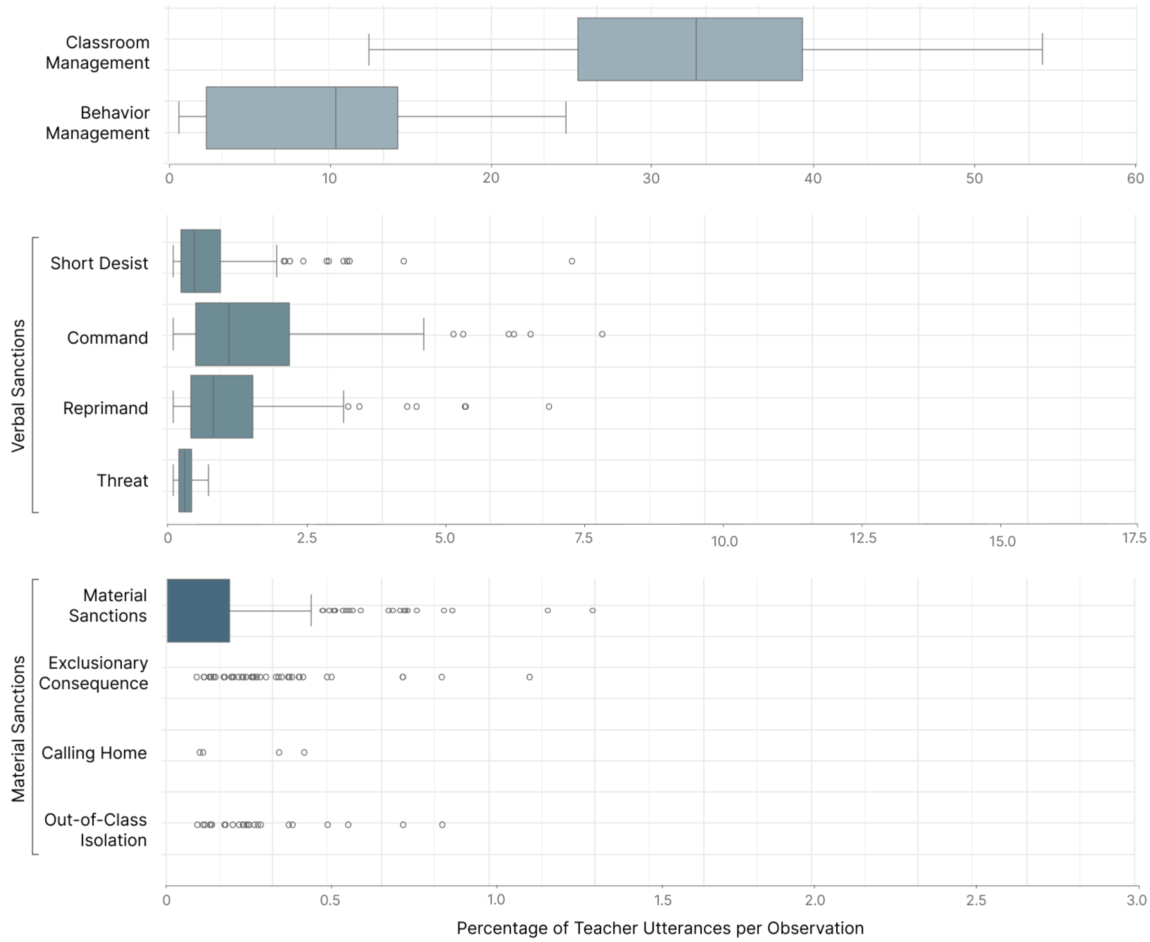
Given the relatively low frequency of specific Behavior Management talk moves, annotating full transcripts became inefficient for these subdimensions. To support more targeted annotation in Phase 2, we used Phase 1 annotations to train a preliminary Behavior Management classifier. This model was applied to the full dataset to identify utterances likely to contain such language. We randomly sampled 200 observations, yielding a filtered sample of 5,720 utterances.

Annotators first coded a shared set of 100 utterances, resolving disagreements through discussion and consensus, followed by an additional 200 utterances independently to assess interrater reliability. The remaining utterances were divided among the six annotators. To increase confidence in the reliability of one dimension—Nonexclusionary Consequences, which showed only moderate IRR—we randomly selected 500 utterances from the remaining sample for duplicate annotation by all six raters. For this dimension, we retained only utterances where all six raters agreed.

Annotators could reject utterances they deemed unrelated to Behavior Management, in which case no subdimension was assigned. The annotations from Phase 2 were also used to augment Phase 1 data for Behavior Management. Specifically, if annotators judged an example was not actually reflective of behavior management (i.e., the preliminary model made an error), this information was used to update the training data for Behavior Management.

Figure C1

Percentage of Classroom Management Talk Moves Among Total Teacher Utterances per Observation in Training Data



Note. The distributions shown here should be interpreted as approximate, intended primarily to illustrate the shape of the training data. Unlike the application dataset, the Classroom Management training data is the only set in which full transcripts (all utterances from an observation) were annotated. Training datasets for other constructs were filtered to reduce the cost of labeling negative examples, using an early version of the Behavior Management model. Consequently, the training sample may not fully represent the frequency of these moves in classroom talk. In constructing this figure, all non-labeled utterances from the relevant transcripts were included and treated as negative examples.

Appendix D:

Modeling and Validation Procedures

We trained independent binary classification models, with each model predicting whether a given utterance matched the definition of a particular talk move, with the exception of verbal sanctions. For the verbal sanctions, we trained a multiclass classification model to predict among four types: Short Desists, Commands, Reprimands, and Threats. Two dimensions—Nonexclusionary Consequences and In-Class Isolation—were not modeled directly, as their values could be inferred more accurately from the predictions of other models. Models were trained for five epochs with a batch size of 8 and a gradient accumulation step of 2. All other parameters were default values.

To address class imbalance, we oversampled minority-class examples through random duplication during training. For extremely rare categories, such as Calling Home and Out-of-Class Isolation, we generated 50 synthetic examples using GPT-4 prompted by high-confidence annotated examples. We use the following prompt: "*Using the following teacher utterances from upper elementary math classroom observations, generate 50 teacher utterances that closely follow the tone, style, and content. Do not add new content not reflected in these utterances.*" We trained RoBERTa-based models rather than relying on GPT models directly for classification, as fine-tuned models offer greater transparency, controllability, and cost-efficiency for large-scale inference.

Table D1 reports performance metrics for all trained models, estimated using 5-fold cross-validation at the utterance level. In this setup, individual utterances are randomly assigned to folds, which is standard since predictions occur at the utterance level. The performance of measures is high across talk moves, with F1 scores ranging between 0.862 and 0.974. Because utterances from the same observation or teacher share context and stylistic features, utterance-level splits may yield somewhat inflated performance estimates. As a robustness check, Table D2 reports results when folds are created at the teacher level (which in most cases aligns with the observation level in the training data, though teachers are not entirely unique in the dataset). In this setup, all utterances from a given teacher are assigned to the same fold, creating variability in the size of training and validation sets across folds. When training sets are smaller, models tend to perform worse; when validation sets are small, performance metrics are more volatile (e.g., folds with zero positive examples). As a result, performance is underestimated and less stable for rarer talk moves. F1 scores remain strong for higher-frequency categories (ranging between 0.815 and 0.905). Although F1 scores are lower for rarer categories (ranging between 0.632 and 0.738), they are still on par with those reported in related work introducing classifiers for teacher talk moves (Alic, 2023; Demszky & Hill, 2023).

Table D1

Model Performance Metrics

Measure	F1	Precision	Recall	N	N+	N Obs.	% Obs.+
Classroom Management	0.887	0.882	0.892	10355	3428	20	100
Behavior Management	0.883	0.846	0.923	8057	4704	20*	100
Verbal Sanction	0.935	0.920	0.949	5720	4657	200	100
Short Desist	0.886	0.855	0.921	4657	755	200	100
Command	0.891	0.879	0.904	4657	2110	200	98
Reprimand	0.862	0.900	0.827	4657	1500	200	93
Threat	0.928	0.921	0.934	4657	291	200	22
Material Sanction	0.908	0.897	0.919	5720	527	200	34
Nonexclusionary Consequence	--	--	--	--	--	--	--
Exclusionary Consequence	0.928	0.926	0.930	5720	360	200	21
Calling Home	0.974	0.950	1.000	5720	98	200	2
In-Class Isolation	--	--	--	--	--	--	--
Out-of-Class Isolation	0.945	0.938	0.952	5720	125	200	12

Note. F1 is an overall indicator of model performance in classification tasks, calculated as the harmonic mean of Precision (how many identified cases are correct) and Recall (how many true cases are identified). N indicates the total number of utterances in the training data, and N+ indicates the total number of positive example utterances in the training data. N Obs. Indicates the number of observations represented in the training data, and % Obs.+ indicates the percentage of those observations with at least one positive example.

Note. The Behavior Management training data includes the 20 observations used in the first phase of annotation, plus individual utterance examples from the second phase that raters relabeled as unrelated to Behavior Management.

Table D2

Model Performance Metrics with Cross Validation Splits at the Teacher-Level

Measure	F1	Precision	Recall
Classroom Management	0.831	0.861	0.804
Behavior Management	0.859	0.813	0.912
Verbal Sanction	0.905	0.887	0.923
Short Desist	0.883	0.860	0.910
Command	0.866	0.857	0.876
Reprimand	0.815	0.835	0.796
Threat	0.896	0.910	0.883
Material Sanction	0.675	0.730	0.627
Nonexclusionary Consequence	--	--	--
Exclusionary Consequence	0.677	0.631	0.730
Calling Home	0.632	0.462	1.000
In-Class Isolation	--	--	--
Out-of-Class Isolation	0.738	0.756	0.721

Appendix E:

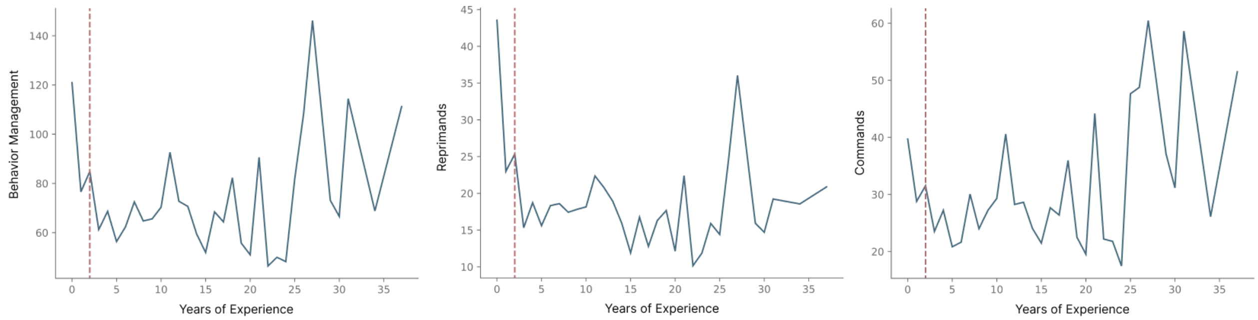
Variance Decomposition (ICC) of Classroom Management Language Measures Across Schools, Teachers, and Instructional Formats

Measure	School	Teacher	Instructional Format	Residual
Classroom Management	1.81	15.53	3.21	79.45
Behavior Management	3.89	17.43	1.47	77.21
Nonbehavioral Classroom Management	0.84	11.06	2.67	85.42
Short Desist	0.53	4.48	0.18	94.81
Command	1.57	15.18	1.11	82.14
Reprimand	3.18	12.32	0.79	83.71
Threat*	5.47	27.76	0.19	66.58
Nonexclusionary Consequence	0.02	1.47	0.20	98.31
Exclusionary Consequence*	10.35	18.04	0.14	71.47
Calling Home*	9.75	38.27	0.00	51.98
In-Class Isolation*	8.72	20.19	0.00	71.09
Out-of-Class Isolation*	10.64	16.66	0.40	72.30

Note. Measures are based on the frequency of observed utterances per classroom observation, except those marked with an asterisk (*), which were binarized to indicate whether any such utterances occurred during the 7.5-minute segment. Nonbehavioral Classroom Management refers to the subset of Classroom Management that is not Behavior Management.

Appendix F:

Relationship between Years of Experience and Frequency of Behavior Management, Commands, and Reprimands.



Note. Y-axes represent the frequency of talk moves in utterances per hour. The dashed vertical line marks the 2-year cutoff for novice teachers.

Appendix G:

Correlations Between Hourly Rate of Each Talk Move, CLASS Observation Scores, and Survey Responses

Measure	Classroom Management	Behavior Management	Short Desist	Command	Reprimand	Threat	Non-exclusionary Consequence	Calling Home	In-Class Isolation	Out-of-Class Isolation	*N*
CLASS Observation Scores											
Behavior management	-0.003*** (0.001)	-0.010*** (0.001)	-0.028*** (0.004)	-0.019*** (0.002)	-0.030*** (0.003)	-0.155*** (0.025)	-0.063*** (0.014)	-0.077 (0.06)	-0.130*** (0.031)	-0.162*** (0.049)	1607
Productivity	-0.004*** (0.001)	-0.009*** (0.001)	-0.025*** (0.005)	-0.016*** (0.002)	-0.025*** (0.003)	-0.149*** (0.026)	-0.075*** (0.017)	-0.050 (0.056)	-0.098*** (0.033)	-0.136** (0.058)	1607
Negative climate	0.003*** (0.001)	0.009*** (0.001)	0.023*** (0.006)	0.014*** (0.003)	0.035*** (0.005)	0.219*** (0.045)	0.096*** (0.021)	0.184** (0.086)	0.133*** (0.049)	0.130** (0.057)	1607
Positive climate	-0.000 (0.000)	-0.004*** (0.001)	-0.015** (0.004)	-0.008*** (0.002)	-0.013*** (0.003)	-0.066*** (0.022)	-0.031** (0.014)	-0.062 (0.05)	-0.060** (0.029)	-0.099** (0.038)	1607
Regard for student perspectives	0.001 (0.001)	-0.002* (0.001)	-0.011*** (0.004)	-0.004** (0.002)	-0.004 (0.003)	-0.029 (0.023)	-0.011 (0.014)	-0.023 (0.042)	0.012 (0.027)	-0.013 (0.044)	1607
Teacher sensitivity	-0.001** (0.000)	-0.006*** (0.001)	-0.022*** (0.004)	-0.013*** (0.002)	-0.016*** (0.003)	-0.094*** (0.023)	-0.038** (0.016)	-0.080 (0.054)	-0.075*** (0.026)	-0.089** (0.041)	1607
Teacher Survey Responses											
Frequency of reprimanding students	0.001** (0.001)	0.003*** (0.001)	0.015*** (0.004)	0.004* (0.003)	0.011*** (0.004)	0.079*** (0.022)	0.032** (0.014)	0.093* (0.054)	0.034 (0.03)	0.073** (0.035)	1559
Frequency of losing time to student	0.001 (0.001)	0.002** (0.001)	0.010** (0.004)	0.003 (0.003)	0.009** (0.004)	0.061** (0.024)	0.010 (0.014)	0.120** (0.047)	0.101*** (0.036)	0.048 (0.041)	1561

misbehavior											
Frequency of feeling disrespected	0.001 (0.001)	0.003** (0.001)	0.009** (0.004)	0.005** (0.002)	0.009** (0.004)	0.098*** (0.025)	0.031** (0.015)	0.063 (0.049)	0.085*** (0.031)	0.073* (0.042)	1561
Student Survey Responses											
My behavior is good	-0.000 (0.001)	-0.003*** (0.001)	-0.011*** (0.004)	-0.005** (0.002)	-0.008*** (0.003)	-0.036** (0.017)	-0.017 (0.012)	-0.010 (0.036)	-0.036 (0.023)	-0.100*** (0.03)	1610
My behavior sometimes annoys the teacher	-0.000 (0.000)	0.002** (0.001)	0.004 (0.003)	0.004** (0.002)	0.007** (0.003)	0.044*** (0.016)	0.023** (0.011)	-0.040 (0.039)	0.036 (0.023)	0.057** (0.026)	1610
My behavior is a problem for the teacher	0.000 (0.000)	0.002** (0.001)	0.004 (0.004)	0.005** (0.002)	0.007** (0.003)	0.035** (0.017)	0.027** (0.014)	-0.044 (0.039)	0.020 (0.022)	0.051* (0.030)	1610

Note. All models include fixed effects for school, school year, and grade level, as well as controls for the total number of utterances and teacher utterances per observation. CLASS and survey measures are standardized. Robust standard errors are clustered at the teacher level. **p* < 0.05. ***p* < 0.01. ****p* < 0.001.