Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1029

A Quantitative Study of Mathematical Language in Upper Elementary Classrooms

Zachary Himmelsbach
Harvard University

Heather C. Hill
Harvard University

Jing Liu
University of Maryland College Park

Dorottya Demszky
Stanford University

This study provides the first large-scale quantitative exploration of mathematical language use in upper elementary U.S. classrooms. Our approach employs natural language processing techniques to describe variation in teachers' and students' use of mathematical language in 1,657 fourth and fifth grade lessons in 317 classrooms in four districts over three years. Students' exposure to mathematical language varies substantially across lessons and between teachers. Results suggest that teacher modeling, defined as the density of mathematical terms in teacher talk, does not substantially cause students to uptake mathematical language, but that teachers may encourage student use of mathematical vocabulary by means other than mere modeling or exposure. However, we also find that teachers who use more mathematical language are more effective at raising student test scores. These findings reveal that teachers who use more mathematical vocabulary are more effective math teachers.

VERSION: August 2024

Suggested citation: Himmelsbach, Zachary, Heather C. Hill, Jing Liu, and Dorottya Demszky. (2024). A Quantitative Study of Mathematical Language in Upper Elementary Classrooms. (EdWorkingPaper: 24-1029). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/1zcm-d071

A Quantitative Study of Mathematical Language in Upper Elementary Classrooms

Zachary Himmelsbach*
Harvard University
Gutman Library, Room 470
6 Appian Way | Cambridge, MA
zah972@g.harvard.edu

Heather C. Hill
Harvard University
Gutman Library, Room 445
6 Appian Way | Cambridge, MA 02138
heather_hill@harvard.edu

Jing Liu
University of Maryland College Park
2203 Benjamin Building
University of Maryland, MD 20740
jliu28@umd.edu

Dorottya Demszky
Stanford University
CERAS Bldg, 520 Galvez Mall
Stanford, CA 94305
ddemszky@stanford.edu

*Corresponding Author

A Quantitative Study of Mathematical Language in Upper Elementary Classrooms

Abstract

This study provides the first large-scale quantitative exploration of mathematical language use in upper elementary U.S. classrooms. Our approach employs natural language processing techniques to describe variation in teachers' and students' use of mathematical language in 1,657 fourth and fifth grade lessons in 317 classrooms in four districts over three years. Students' exposure to mathematical language varies substantially across lessons and between teachers. Results suggest that teacher modeling, defined as the density of mathematical terms in teacher talk, does not substantially cause students to uptake mathematical language, but that teachers may encourage student use of mathematical vocabulary by means other than mere modeling or exposure. However, we also find that teachers who use more mathematical language are more effective at raising student test scores. These findings reveal that teachers who use more mathematical vocabulary are more effective math teachers.

Keywords: teacher research, mathematics education, student development, vocabulary, language comprehension/development, instructional practices, classroom research, achievement

A Quantitative Study of Mathematical Language in Upper Elementary Classrooms

In contrast with everyday speech and writing, mathematical language uses a specialized vocabulary with precise meanings (Pimm, 2019; Schleppegrell, 2007). This precise language serves three functions in a classroom setting: it is a medium of communication, a foundation of students' understanding, and a tool used to assess students' comprehension (Thompson & Rubenstein, 2000). Given these multiple functions, teachers seeking to encourage student use of mathematical language typically face several tasks. At the most basic level, teachers connect mathematical terms to the mathematical content and representations that those terms signify and to students' own intuitive, informal understandings of that content, effectively defining terms for their students (Rubenstein & Thompson, 2002). Teachers also serve as the 'native speaker' of mathematics, modeling the fluent use of mathematical terms in their instruction (Pimm, 2019). Teachers can additionally encourage students to practice speaking about mathematics using the appropriate specialized terms (Hughes, Powell, & Stevens 2016).

Teachers vary in their engagement with each of these teaching tasks (Hughes, Powell, & Stevens 2016; Lane, O'Meara & Walsh, 2019; Ernst-Slavit & Mason, 2011). Some teachers intentionally populate their lessons with mathematical language in order to model “speaking mathematically” (Khisty & Chval, 2002) while others prefer to use more colloquial language in place of mathematical terms (Hughes, Powell & Stevens 2016). Teachers also vary in the extent to which they employ explicit strategies to elicit mathematical vocabulary from their class, for instance by repeatedly linking mathematical terms to students’ informal terminology, (see, e.g., Echevarria et al., 2017; Khisty & Chval, 2002; Rubenstein & Thompson, 2002). Teachers’ use of mathematical terms and their encouragement of students’ mathematical vocabulary use may be supported by their mathematical knowledge and prior learning experiences, but it can also be limited by their own lack of fluency with mathematical terms (e.g., Gürefe, 2018) or support for mathematical vocabulary in their curriculum (Barnes & Stephens, 2019).

To date, studies of classroom mathematical language and its effects on student learning have been mainly limited to theoretical inquiry, case studies, and a few small program evaluations. Using techniques from natural language processing (NLP), we undertake a large-scale study of teachers’ and students’ use of mathematical vocabulary. Our analyses address the following sets of questions:

RQ 1: How frequently do upper-elementary teachers and students in this sample use mathematical vocabulary? How much does this vary across lessons?

RQ 2: Do teachers differ systematically in the amount of mathematical vocabulary to which they expose their students? Do their classrooms systematically differ in how much mathematical vocabulary students use? Are such differences explained by teachers' background or mathematical knowledge?

RQ 3: Does teacher use of mathematical terms (modeling) affect students' use of mathematical terms and their standardized test score outcomes? Is there evidence that teachers' encouragement of student vocabulary use relates causally to student language use and their later test score outcomes?

We answer these questions using anonymized transcripts from 1,657 4th-5th grade math lessons taught by 317 teachers over three years. The National Center for Teacher Effectiveness (NCTE) collected these data between 2010-2013 across four school districts that largely serve historically marginalized students (Demszky & Hill, 2023). Using NLP techniques, we create a dictionary of mathematical terms and extract their occurrences from transcripts of student and teacher speech. We then use this data to generate measures, separately for teachers and students, of mathematical vocabulary use in lessons. We combine these measures with NCTE data on teachers' mathematical coursework, assessment of their mathematical knowledge for teaching (MKT), and their students' standardized test scores. We generate causal estimates of the impact of vocabulary use on student achievement using the random assignment of students to teachers in the NCTE study design; in the first two years of the study, students were assigned to teachers as per typical school policy, but, in the third year, students were randomly assigned to teachers. We leverage pre-randomization measures of teacher and student vocabulary use to test whether being assigned to a teacher who used more mathematical vocabulary - or whose past students used more mathematical vocabulary - affects student test scores.

Background

Teachers may use a range of approaches to promote students' mathematical language development. Some simply model fluent use of mathematical language; Khisty and Chval (2002), for instance, describe a teacher who encourages student mathematical talk mainly through modeling it herself, using precise vocabulary and connecting terms to their meanings. Other teachers may employ a variety of more explicit instructional strategies to encourage students' use of mathematical vocabulary. For instance, teachers may post mathematical terms on a "word wall" to help make new words meaningful, modulate their speech to support meaning, support student talk about mathematics problems, ask them to write responses that feature mathematical terms, and teach about the etymology of mathematical terms (Rubenstein & Thompson, 2002; Thompson & Rubenstein, 2000). Teachers may also, quite simply, press students for the use of precise mathematical language during mathematical discussions.

Teachers' approaches to mathematical language may be influenced by several factors. They may hold differing beliefs about whether mathematical vocabulary supports or frustrates learning of the targeted mathematical concepts: some teachers populate their lessons with mathematical language to demonstrate the utility of being able to "speak mathematically" (Khisty & Chval, 2002) while others prefer to use colloquial language in place of mathematical terms, believing it enhances students' understanding (Hughes, Powell & Stevens 2016). Another factor may be teachers' own mathematical knowledge and the degree to which they are fluent in mathematical speech themselves (e.g., Gürefe, 2018). Teachers' modeling and encouragement of the use of mathematical terms may also result from their training or experience as a teacher, or by the extent to which their curriculum prioritizes the use of and teaching about mathematical vocabulary (Barnes & Stephens, 2019).

Given the range in potential approaches to mathematical language in classrooms - and known variability in teachers' mathematical knowledge, beliefs, and use of curriculum materials - it is not surprising that the use of mathematical vocabulary varies markedly by lesson and teacher (Khisty & Chval, 2002; Hughes, Powell & Stevens, 2016). However, the magnitude of this variation remains unknown, as most investigations of this phenomena remain small in scale (e.g. Ernst-Slavit & Mason, 2011; Pimm, 2019; Schleppegrell, 2007), likely because the costs of traditional, observation-based methods prohibit the analysis of large samples of lessons. The same constraint applies to studies that compare teacher and students' classroom use of mathematical vocabulary with later student achievement. While the correlation between formal tests assessing student mathematical vocabulary and tests assessing mathematical problem-solving performance has been well-documented (see Lin, Peng, & Zeng, 2021 for a meta-analysis of related studies), to our knowledge no classroom studies exist, leaving the field unsure of how a focus on mathematical vocabulary might aid students' overall learning.

The existing literature documenting mathematical vocabulary use in classrooms presents two additional limitations. First, it does not consider teachers' modeling of mathematical vocabulary and their encouragement of students' vocabulary use as separate phenomena. It may be the case that teachers' modeling has negligible effects on students' mathematical vocabulary use and that other approaches to encouraging vocabulary use are more effective. Or it may be the case that teacher modeling itself drives student vocabulary acquisition, even when other strategies to encourage student vocabulary use are not employed.

Secondly, classroom studies of mathematical vocabulary and achievement have been almost exclusively correlational. Empirical evidence for a causal link between students' mathematical vocabulary and their outcomes is limited to a few extant studies focused on explicit teaching of mathematical terms directly to students, rather than testing the broader array of pathways through which classroom mathematical vocabulary may influence outcomes. For instance, one experimental study randomly assigned a small group of preschool students to a mathematical language-improving intervention and found positive effects on math performance (Purpura et al, 2017; see also. Hassinger-Das et al., 2015; Powell & Driver, 2014). Yet no experiment tests whether and how different approaches to mathematical vocabulary use in classrooms increase students' own use of mathematical vocabulary, or, by extension their performance on standardized assessments.

NLP techniques present an opportunity to describe mathematical language use and test for connections between classroom use of mathematical vocabulary and student outcomes. Compared to human-based classroom observations, these lower-cost data processing techniques allow scholars to measure teaching behaviors at a wider scale and have been used to study a range of other classroom behaviors. For instance, scholars have measured the cognitive demand and goal-specificity of teachers' instructions to their students (Dale et al., 2022), whether teachers' questions are content-specific (Stone et al, 2019) or authentic (Kelly et al, 2018), and teachers' uptake of students' ideas (Demszky et al., 2021). Studies connecting NLP measures of teaching practice with student outcomes are rarer. However, Demszky and Hill (2023) and Liu and Cohen (2021) generate measures for a range of teaching practices and explore their association with students' test scores. We follow this approach to investigate teachers' and students' use of mathematical vocabulary and its relationship to students' outcomes.

Data and Methods

Data

The NCTE Transcript Data (Demszky & Hill, 2023) contain transcriptions of student and teacher speech from 1,657 fourth and fifth grade mathematics lessons in US public schools. The lessons were recorded in four large school districts between 2010 and 2013. For the first two years of the study, students were assigned to teachers following the schools' and districts' typical procedures. In the final year of the study, school staff composed class lists and then the NCTE study team randomly assigned teachers to those classes.

The students represented in the dataset disproportionately belong to historically marginalized populations. District records indicate that 43% of students in the sample are Black, and 23% are Hispanic. Of the remainder, most are white (23%). A majority qualified for free or reduced-priced lunch (67%), and 21% were reported in district records as having limited English proficiency. Students were approximately balanced on gender. In contrast, the majority of the 317 teachers in the sample are white (65%), and the vast majority are female (84%). Such disparities between student and teacher demographics are typical of the US education context. Teachers in the sample had, on average, 10 years of experience, but this varied substantially (SD=7.3). Only 6% held bachelor's or graduate degrees in mathematics.

Lesson audio was captured from two sources: lapel microphones worn by teachers and two bidirectional microphones mounted on cameras stationed in the middle of the classroom. Professional transcribers under contract to a commercial transcription company converted the audio to text. Each transcript was fully anonymized, with indicators like “Student J” and “Teacher” replacing student and teacher names, respectively. Unintelligible words, meaning words not discernible due to classroom noise, low-volume student voices, or other sound issues, were transcribed as [Inaudible]. When transcribers were unsure of a word, they transcribed it within brackets, e.g. “I like how you drew that [figure].” Transcriptionists also used square brackets for descriptions such as [crosstalk] and [laughter]. A team at Harvard later worked to clean up the transcripts, using high-quality headphones, and replaced many [inaudible] marks with teacher or student speech. Almost all teacher talk and the majority of student talk could be transcribed: only 4% of teacher utterances and 21% of student utterances contained one or more [inaudible] words.

The average transcript contains 5,733 words. Of these, 88% were spoken by teachers. Each teacher in the study was recorded multiple times, typically three times per year. Due to teacher turnover in the districts and within the study, 189 of the teachers were recorded for two years. Our experimental sample comprises the 100 teachers who were recorded in both the randomization year and the prior year(s).

Measure construction

To create an initial dictionary of mathematical terms, we scraped publicly available glossaries from two K-6 math curricula, Zearn and Illustrated Mathematics. A small number of terms – “row” for example – were dropped because they have common, non-mathematical meanings in classrooms. Some terms were stemmed to match multiple forms of the word (e.g. “multiplication” is stemmed to “multipl” to also match “multiply” or “multiples”). Initialisms, like “LCM”, were supplemented with their expanded forms. So, for example, both “LCM” and “least common multiple” are included. In the end, we generated a list of 256 mathematical terms. Of these, 224 terms appear at least once in the NCTE transcript data. Many of the words that never appear are statistics terms from the 6th grade curriculum glossaries. Our transcripts come from 4th and 5th grade classrooms, so this makes sense.

To estimate how many terms we might have missed, we audited a random sample of teacher utterances from the transcripts. We stratified our sample on whether the utterance contained at least one mathematical term in our list. The sample included 2,000 randomly drawn utterances (out of about 580,000 total in the corpus), with half of them containing at least one mathematical term from our dictionary. The stratification was motivated by the idea that utterances with some mathematical terms might be more likely to contain additional ones, because we know those utterances are math-related (as opposed to being focused on classroom management). A manual check of the randomly selected utterances revealed a single mathematical term in this sample that was not included in our list. This term was not used in any other utterance in the dataset. Given the results of this audit, we estimate that we have missed a very small number of mathematical terms spoken in the sample. In addition, any terms we have missed are likely to be rarely used.

Our final list of 224 terms spans elementary operations (e.g. “multiply”, “divide”, “add”, “subtract”, “product”, “sum”), geometric terms (e.g. “angle”, “rectangle”, “graph”, “polygon”, “quadrilateral”), measurement-related terms (e.g. “units”, “measure”, “length”, “meter”), number-related terms (e.g. “digit”, “remainder”, “number line”, “factors”, “place value”), and others (e.g. “median”, “algorithm”, “ordered pair”). These terms were stemmed to capture alternate forms that share the same root, such as “subtract” and “subtraction”. With this list, we calculated a count of the number of unique mathematical terms in each transcribed utterance. This choice means that repeating the same term multiple times within an utterance will have the same weight in our measures as using it a single time.

From these utterance-level counts, we constructed two lesson-level measures of mathematical language. One captures teachers' modeling of mathematical terms in their own speech and is a sum of the utterance-level counts of mathematical terms in a particular lesson. The second is a measure of students' mathematical term use, which similarly is a sum of their utterance-level counts in each lesson. We use this second measure as a proxy for teachers' encouragement of student mathematical vocabulary use, either through modeling alone or through the strategies listed above. To produce teacher-level measures of modeling and encouragement, we averaged their lesson-level measures. We present additional information on the construction of these measures in Appendix A.

Our measures imply a broad definition of mathematical language use. We made no prescriptive choices regarding what math vocabulary terms teachers and students should be using. As we discuss in the conclusion of the paper, future work should explore variation in the complexity of classroom mathematical language and differences in how vocabulary terms are used.

When we analyze data from the randomization year, we treat teachers' modeling and encouragement measures from the pre-randomization years as covariates and their students' amount of mathematical vocabulary after random assignment to teachers as the outcome. We note that the encouragement measure can only be considered a proxy as student mathematical vocabulary use will be affected by not only teacher encouragement but also by the students' prior knowledge and experience. One benefit of the randomization in this study is that, on average, classrooms within school-grades have similar compositions of student backgrounds, making observed differences in student vocabulary use attributable to differences between teachers.

The NCTE data also include several covariates that potentially explain differences in teachers' use of mathematical vocabulary. These include instructors' mathematical knowledge for teaching (MKT; alpha = 0.87), a measure of the specialized knowledge held by teachers and thought to enable the teaching of content to children (Ball, Thames & Phelps, 2008). We hypothesize that teachers' MKT may support teachers' and their students' use of mathematical vocabulary in the classroom. A second set of covariates proxied for teacher background: the number of mathematics content and mathematics education courses they reported taking, and their years of experience.

Finally, the NCTE data contains students' state standardized test scores for each year of the study and their demographic information, provided by the school districts.

Analysis

To answer our first research question - how frequently our sample of upper-elementary teachers and students use mathematical terms and how much this varies across lessons - we present sample medians and sample standard deviations of our mathematical language measures for teacher and student speech. To answer our second question, we estimate the degree to which teachers systematically vary in their use of mathematical terms by fitting a three-level hierarchical model and comparing the estimated variance components at the teacher, school, and lesson level:

$$Math\_Terms_{lisd} = \mu_i + \zeta_s + \lambda_d + \varepsilon_{lisd}$$

$$\text{Where } \mu_i \sim N(0, \sigma_i^2)$$

$$\text{and } \zeta_s \sim N(0, \sigma_s^2)$$

where lessons (l) are nested within teachers (i), and teachers are nested within schools (s). We estimate this model with data at the teacher-lesson level. We include fixed effects λ_d for the districts to adjust for differences in curricular materials and related instructional guidance. However, estimates of the percentage of total variance accounted for by each level differ by less than 1.5 percentage points between the models with and without district fixed effects. We take this multi-level approach because we observe only a sample of each teachers' lessons, and lesson-level features - such as the topic of the lesson - are likely to drive differences in teachers' observed average use of mathematical terms per lesson. Therefore, using only the observed teacher averages would lead to overestimating the differences between teachers' mathematical language use. To address this concern, we calculate empirical Bayes shrunken estimates of the teacher averages based on the above model and report their interquartile range. We estimate the above models for both the teachers' speech and, separately, for the speech of their students.

To test whether differences in teachers' use of mathematical language are explained by differences in their mathematical knowledge for teaching or their observed background, we fit regression models to test for associations between our mathematical language measures and observable characteristics.

$$Math\_Terms_{isdt} = Characteristic_{isdt} + \varepsilon_{isdt}$$

We estimate the linear relationship for each observed characteristic at the teacher-by-year level separately and in a joint model. In this model, the outcome is the average number of mathematical terms per lesson across all a teachers' observed lessons in year *t*. Teachers with missing variables on these characteristics are excluded from this analysis. Standard errors are clustered at the teacher level. We repeat these analyses with the amount of mathematical language used by the teachers' students as the outcome. We also test whether including these teacher-level covariates explains the estimated teacher level variance from our hierarchical model.

Finally, we take advantage of the random assignment of students to teachers in the third year of the NCTE study to estimate the causal impacts of being assigned to a teacher who models and/or encourages more mathematical vocabulary use in their lessons. For each measure (teacher modeling and teacher encouragement) we estimate effects on two outcomes: students' mathematical vocabulary use and test scores. Additionally, to overcome potential bias caused by the reflection problem discussed in the peer effects literature (i.e., teachers' and students' use of mathematical terms in the same classroom are functions of each other) (Manski, 1993), we adjust for lagged measures of modeling and encouragement in our models.

$$Y_{igs} = \alpha_0 + \alpha_1 Past\_Teacher\_Math\_Vocab_{igs} + \alpha_2 Past\_Student\_Math\_Vocab_{igs} + \lambda_{gs} + \beta X + \varepsilon_{igs}$$

In models of student vocabulary use, Y_{igs} is the average number of math terms - per lesson - used by students in teacher i's classes, *in the randomization year*. In models of test scores, Y_{igs} is a student's standardized end-of-year test score. For the past student and teacher vocabulary use measures, we take the averages over time if we observe two pre-randomization years. We include school-by-grade fixed effects, λ_{sg}, to account for the random assignment (of students to teachers) occurring within school-grades. This allows us to estimate the causal effect on students' spoken vocabulary and test scores of being assigned to a teacher who, prior to randomization, modeled more mathematical vocabulary or whose students used more mathematical vocabulary. If teachers play no role in student use of mathematical language - i.e. the pre-randomization correlation between student and teacher mathematical language use is entirely driven by the sorting of more mathematically fluent students to more mathematically fluent teachers - we expect these effects, α_1 and α_2, to be indistinguishable from zero. In models where the outcome is test scores, we also include student-level covariates, X, including prior standardized test scores, English language learner status, free and reduced priced lunch status, and fixed effects for demographics (gender and race/ethnicity).

[Page Break]

Results

Teachers' and students' use of mathematical terms across lessons

Figure 1: Teacher and Student Use of Mathematical Vocabulary

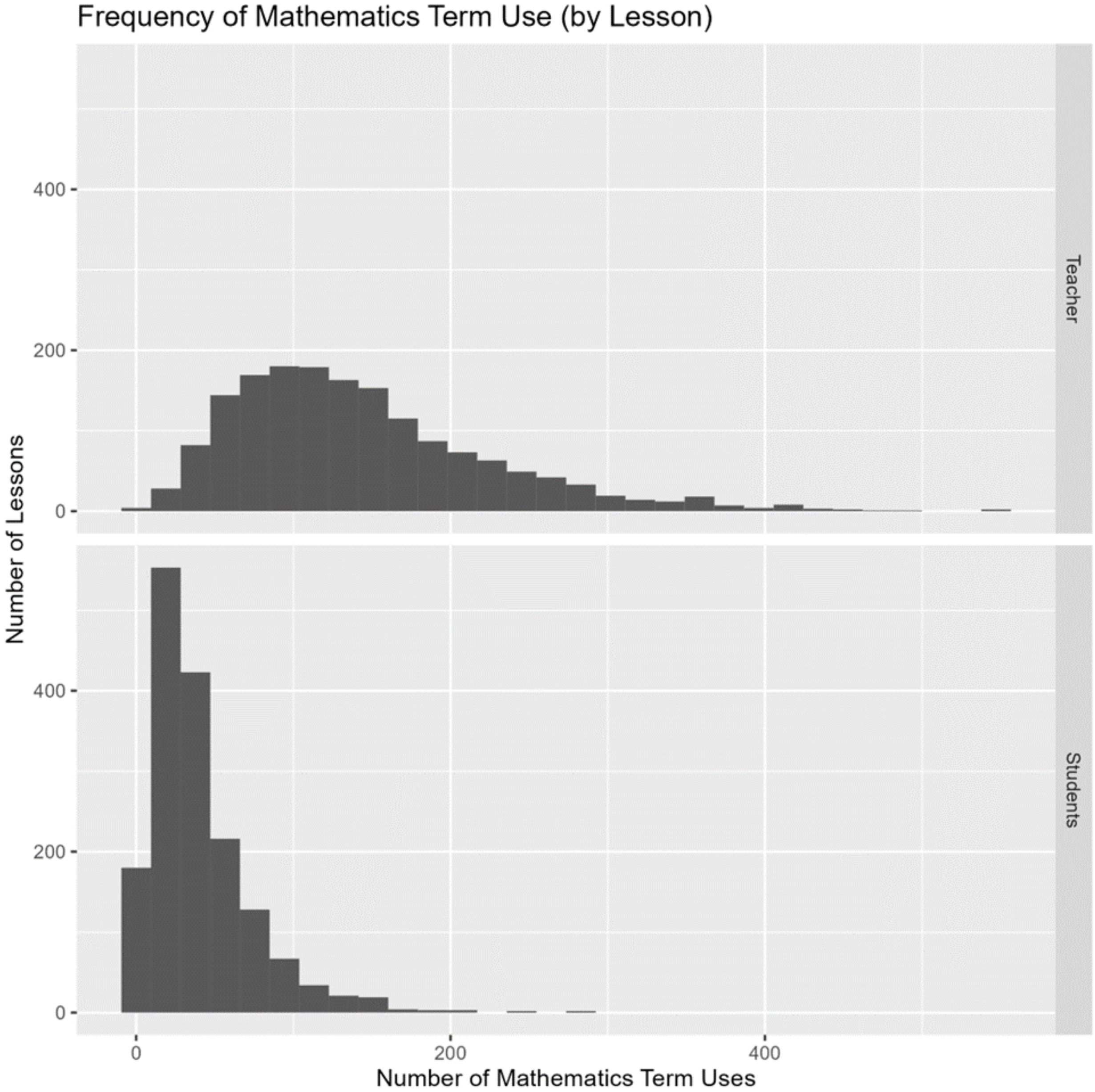

Figure 1 shows that, in the median lesson across all teachers and years in our sample, teachers modeled the use of mathematical terms 127 times, but this frequency varied substantially across lessons (sd = 82), precipitating large differences in students' exposure to mathematical vocabulary. Lessons in the 25th percentile feature 84 terms, while lessons at the 75th percentile of this measure feature 185 terms.

Lessons also differed in the amount of mathematical vocabulary used by students. In the median lesson, a teacher heard 38 uses of mathematical terms from their students. As above, this varies, with lessons at the 25th and 75th percentiles eliciting 26 and 50 math terms, respectively. Mathematical terms make up a greater proportion of student speech (5.7% vs teachers' 2.8%). Lesson-level measures of teacher and student use of mathematical vocabulary correlate moderately ($r = .65$).

Variation in the use of mathematical terms among teachers and their students

Next, we examined the extent to which teachers in our sample varied systematically in their frequency of modeling mathematical terms. The average teacher modeled mathematics vocabulary 140 times per lesson (sd=54). Variance decomposition reveals persistent teacher differences, with teachers accounting for

14% of the variance in the mathematical vocabulary observed at the lesson level ($p < 0.01$). The high level of residual variance (86%) indicates differences within teachers (e.g., based on the content of lessons). Empirical Bayes estimates suggest that the gap between teachers at the 25th and 75th percentile is approximately 28 terms per lesson. This represents a difference of 4,480 exposures to mathematical terms over the course of the school year. For students' use of mathematical vocabulary, the results are nearly identical, with 15% of the variation in the number of mathematics terms used by students accounted for at the teacher level. Empirical Bayes estimates suggest that the gap between classrooms at the 25th and 75th percentile is approximately 10 student terms per lesson. This represents a difference of 1,800 student-spoken mathematical terms over the course of the school year.

Of the observed teacher characteristics, a joint regression model finds that only MKT assessment scores significantly predict more frequent teacher use of mathematical vocabulary (see table 1A). Having an MKT score 1 SD higher is associated with using 9 more mathematical terms per lesson, on average. This represents 8% of the mathematical vocabulary used in the median lesson. Estimates of the effects of teachers' number of college mathematics courses taken and their number of mathematics methods courses were insignificant, and the confidence intervals are precise enough to rule out strong associations. Although this analysis suggests meaningful variation among teachers with different levels of MKT, the teacher-level variance in mathematical vocabulary remains unchanged by the teacher characteristics we observe. None of the observed teacher covariates significantly explain differences in students' use of mathematical vocabulary (see table 1B). When we add these teacher-level covariates to our models predicting student mathematical vocabulary, the estimated teacher-level variances are not reduced.

Table 1A: Teacher Characteristics and Mathematical Vocabulary Use

	Dependent variable: Math Vocabulary Used by Teacher (per lesson)	
	Bivariate Models	Joint Model
	(1)	(2)
MKT Score (SDs)	9.93^{***}	8.63^{**}
	(2.30)	(3.49)
# Methods Courses	-5.32	-5.41
	(3.80)	(3.95)
# Math Courses	-0.05^{*}	-0.05
	(0.03)	(0.03)
Experience	-0.22	-0.03
	(0.43)	(0.45)
Constant	0.00	164.77^{***}
	(0.00)	(10.28)
R-squared	-	0.01
F Statistic (df = 4; 832)	-	2.99**
Observations	825	825
Note:		*p<0.1**p<0.05***p<0.01

Caption: The first column shows coefficient estimates from separate bivariate regressions. For example, the first reported coefficient in column 1 is from a simple regression of a teachers' number of math terms per lesson (in a given year) on their MKT score (with no additional controls). The second column shows estimates from the joint model. Standard errors are clustered at the teacher level.

Table 1B: Teacher Characteristics and Student Mathematical Vocabulary Use

	Dependent variable: Students' Use of Mathematical Vocabulary (per lesson)	
	Bivariate Models	Joint Model
	(1)	(2)
MKT Score (SDs)	2.02**	2.10
	(0.96)	(1.43)
# Content Courses	-0.41	-1.19
	(1.56)	(1.61)
# Math Courses	-0.01	-0.01
	(0.01)	(0.01)
Experience	0.15	0.20
	(0.18)	(0.18)
Constant		44.99***
		(4.20)
R-squared	-	0.01
F Statistic (df = 4; 832)	-	2.99**
Observations	825	825
Note:		*p<0.1**p<0.05***p<0.01

Caption: The first column shows coefficient estimates from separate bivariate regressions. For example, the first reported coefficient in column 1 is from a simple regression of students' number of math terms per lesson (in a given year) on their teacher's MKT score (with no additional controls). The second column shows estimates from the joint model. Standard errors are clustered at the teacher level.

Connection between teachers' use of mathematical terms and student outcomes

To explore the causal relationship between teachers' use of mathematical terms and student outcomes, we begin with students' use of mathematical vocabulary in the randomization year (Table 2). We find that when entered in the model singly, both teacher (1) and student use of mathematical terms (2) in the pre-randomization year predict student vocabulary use in the randomization year. However, the effect of teachers' use of terms in the pre-randomization year is not as strong and drops toward zero in the joint model; meanwhile, the impact of pre-randomization students use of terms remains roughly the same size in the joint model, though becomes insignificant (3). This suggests that some elements of teacher practice effectively encourage more mathematical vocabulary use from students. It additionally demonstrates that differences in classes' use of mathematical vocabulary is not attributable solely to selective sorting.

Table 2: Effects on Student Vocabulary Use

	Dependent variable:		
	Post-randomization Student Vocabulary (per Lesson)		
	(1)	(2)	(3)
Pre-rand Teacher Terms per Lesson	0.09^{*}		0.03
	(0.05)		(0.06)
Pre-rand Student Terms per Lesson		0.29^{**}	0.25
		(0.12)	(0.15)
Constant	11.19	14.89	12.91
	(18.12)	(17.18)	(17.87)
School-by-Grade Fixed-Effects	Yes	Yes	Yes
Observations	100	100	100
R^2	0.53	0.55	0.55
F Statistic	1.54^{*}	1.68^{**}	1.62^{**}
Note:		$^{*}p<0.1^{**}p<0.05^{***}p<0.01$	

Finally, we turn to student test score outcomes. We find that teachers who used more mathematical vocabulary prior to randomization caused higher test scores in the randomization year. Having a teacher who exposed their past students to 1 SD more mathematical terms per lesson improved standardized test scores by 0.06 standard deviations. While we cannot reject that the effects are the same for these two measures, the F-test of the joint model suggests that as a pair, teachers' and their students' use of mathematical language predict which teachers are causally more effective at raising student test scores. Appendix B presents results of our testing on experimental balance, including evidence of imbalances in some of the randomization blocks. The models presented here adjust for those differences.

[Page Break]

Table 3: Student Test Score Effects

	Dependent variable:	
	Standardized State Math Test Score	Standardized State Math Test Score
	(1)	(2)
Pre-randomization Teacher Modeling	0.06^{***}	
	(0.02)	
Pre-randomization Student Math Vocabulary Use		0.03
		(0.02)
Constant	-0.18^{**}	-0.21^{**}
	(0.09)	(0.09)
Block FEs	Yes	Yes
Student Covariates	Yes	Yes
Observations	2,201	2,201
R^2	0.61	0.61
Residual Std. Error (df = 2149)	0.60	0.61
F Statistic (df = 51; 2149)	65.21^{***}	64.85^{***}
Note:		*p<0.1**p<0.05***p<0.01

Caption: Student-level covariates include student gender, special education status, race/ethnicity, free-reduced price lunch status, English language learner status, and prior standardized state math test scores.

Discussion

We observed that teacher and student mathematical vocabulary varies significantly across lessons. Decomposing math language scores by lesson and teacher suggests that a significant portion of this variability (14%) lies at the teacher level. Students of teachers in the 75th percentile would hear 28 more mathematical terms per lesson (4,480 per year) than students of a teacher in the 25th percentile. Likewise, they would speak 10 more terms than students in the 25th percentile. At both the lesson and teacher level, this variability mirrors differences we see in instruction generally (Hill, Litke, Lynch 2018).

Observed teacher characteristics explain very little of the variation described above: Teachers with higher MKT scores do use more mathematical vocabulary per lesson, on average, but differences in teachers' use of mathematical language are not meaningfully explained by this or other teacher characteristics. No observed teacher characteristic significantly predicted differences in their students' use of mathematical vocabulary. The weak explanatory power of teacher characteristics in predicting mathematical language use suggests that other factors may be at play, including variation across curriculum in mathematical language (Barnes & Stephens, 2019), teachers' beliefs about the appropriateness of mathematical terms for their students (Hughes, Powell & Stevens 2016), and variability in lesson formats (e.g., groupwork) that could impact teachers' overall amount of classroom talk and thus mathematical term use.

Students randomly assigned to teachers' who used more mathematical vocabulary in their previous classrooms scored higher on standardized tests of mathematics. This implies that teachers who expose their students to more mathematical vocabulary are more effective teachers of mathematics. Interestingly, across value-added studies, a teacher one standard deviation above the mean in effectiveness raises math scores by between .10 and .15 (Bacher-Hicks & Koedel, 2023); our estimate of the effect of being assigned to a teacher who uses one standard deviation more mathematical language accounts for roughly half of this variation.

Importantly, students were randomly assigned to teachers with these qualities, not to particular mathematical language-focused interventions. The measures we have constructed may be correlated with other teacher effective practices. For example, it may be the case that teachers who use more mathematical vocabulary also happen to be teachers who demonstrate more "worked examples" in class. In that case, it may be the worked examples that cause higher test scores, rather than differences in mathematical language use. Hence, this study does not indicate that teachers modeling - or encouraging more mathematical vocabulary - causes higher test scores, only that the teachers who do these things are, on average, more effective at improving student test performance. It is likely that teacher and student language use belong to a constellation of instructional elements that together impact student outcomes. That said, evidence that a math language-improving program impacts overall student outcomes (Purpura et al, 2017) suggests that mathematical language may be an important member of that constellation.

This study is among the first to use NLP techniques to describe and analyze instructional quality in a large sample of U.S. classrooms. Such techniques are increasingly deployed to measure key aspects of instructional quality and have the advantage of analyzing large amounts of data at very low cost. One natural extension of such measures might be their use in providing feedback to teachers. Drawing attention to the density of teachers' use of mathematical terms and opportunities for students to practice using such terms may prove effective in increasing the prominence of mathematical language in teachers' speech.

In addition to exploring applications of these measures, we see two other areas for future research. One is grounded in the fact that we did not measure the quality of teacher use of mathematical vocabulary. Teachers may use terms incorrectly, e.g., mis-defining terms or using them in non-standard ways. Our automated measure cannot capture such uses, but datasets tagged with such events, such as the one here, might be used to train models to detect this issue. As well, our measure cannot capture variation in the instructional strategies teachers employ to improve the mathematical language of their students. Because some approaches may yield greater improvements of students' use of mathematical language, future research should construct NLP-based measures of such specific strategies to describe their prevalence and their relationship to measures of students' outcomes.

References

Bacher-Hicks, A., Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. In Handbook of the Economics of Education. Elsevier Science & Technology.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special?.

Barnes, E. M., & Stephens, S. J. (2019). Supporting mathematics vocabulary instruction through mathematics curricula. The Curriculum Journal, 30(3), 322-341.

Dale, M. E., Godley, A. J., Capello, S. A., Donnelly, P. J., D'Mello, S. K., & Kelly, S. P. (2022). Toward the automated analysis of teacher talk in secondary ELA classrooms. Teaching and Teacher Education, 110, 103584.

Demszky, D., & Hill, H. (2023). The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. In 18th Workshop on Innovative Use of NLP for Building Educational Applications.

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. 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), 1638–1653. Association for Computational Linguistics.

Ernst-Slavit, G., & Mason, M. R. (2011). "Words that hold us up:" Teacher talk and academic language in five upper elementary classrooms. Linguistics and Education, 22(4), 430-440.

Gürefe, N. (2018). Mathematical Language Skills of Mathematics Prospective Teachers. Universal Journal of Educational Research, 6(4), 661-671.

Hassinger-Das, B., Jordan, N. C., & Dyson, N. (2015). Reading Stories to Learn Math. *The Elementary School Journal*, *116*(2), 242–264. https://doi.org/10.1086/683986

Hill, H. C., Litke, E., & Lynch, K. (2018). Learning lessons from instruction: Descriptive results from an observational study of urban elementary classrooms. Teachers College Record, 120(12), 1-46.

Hughes, E. M., Powell, S. R., & Stevens, E. A. (2016). Supporting clear and concise mathematics language: Instead of that, say this. Teaching Exceptional Children, 49(1), 7-17.

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.

Khisty, L. L., & Chval, K. B. (2002). Pedagogic discourse and equity in mathematics: When teachers’ talk matters. Mathematics education research journal, 14(3), 154-168.

Lane, C., O'Meara, N., & Walsh, R. (2019). Pre-service mathematics teachers' use of the mathematics register. Issues in Educational Research, 29(3), 790-806.

Lin, X., Peng, P., & Zeng, J. (2021). Understanding the Relation between Mathematics Vocabulary and Mathematics Performance A Meta-analysis. The Elementary School Journal, 121(3), 504–540. https://doi.org/10.1086/712504

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.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. The review of economic studies, 60(3), 531-542.

Pimm, D. (2019). Routledge Revivals: Speaking Mathematically (1987): Communication in Mathematics Classrooms (Vol. 4). Routledge.

Powell, S. R., & Driver, M. K. (2015). The Influence of Mathematics Vocabulary Instruction Embedded Within Addition Tutoring for First-Grade Students With Mathematics Difficulty. *Learning Disability Quarterly*, *38*(4), 221–233. https://doi.org/10.1177/0731948714564574

Purpura, D. J., Napoli, A. R., Wehrspann, E. A., & Gold, Z. S. (2017). Causal Connections Between Mathematical Language and Mathematical Knowledge: A

Dialogic Reading Intervention. Journal of Research on Educational Effectiveness, 10(1), 116–137. https://doi.org/10.1080/19345747.2016.1204639

Rubenstein, R. N., & Thompson, D. R. (2002). Understanding and Supporting Children's Mathematical Vocabulary Development. Teaching Children Mathematics, 9(2), 107–112. https://doi.org/10.5951/TCM.9.2.0107

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. Reading & writing quarterly, 23(2), 139-159.

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.

Thompson, D. R., & Rubenstein, R. N. (2000). Learning mathematics vocabulary: Potential pitfalls and instructional strategies. The Mathematics Teacher, 93(7), 568-574.

Supplemental Materials:

Appendix A: Methodology

We took several steps to ensure our list represents the mathematical vocabulary used in 4th and 5th grade classrooms. To begin, we generated word2vec embeddings for all words in our corpus of transcripts. These embeddings represent each word as a vector of 50 numeric values. Words that are nearby in the embedding space tend to be used in similar contexts and therefore have related meanings. To fit embeddings we employed a skip-gram training method, wherein the embedding of a word is chosen to maximize its power to predict the words that appear around it. The skip-gram training objective has been found to outperform the primary alternative approach, known as continuous bag of words, for smaller corpora such as ours. We trained the embeddings in a shallow neural network using the word2vec package (Wijffels, 2021). This method results in static embeddings where each token is represented by the same vector for all its appearances in the transcripts (e.g. the word “square” is always represented by the same set of numeric values). Transformer models, a deeper neural network architecture, would allow for contextual embeddings that vary across usages, but given that our target mathematical terms are not polysemous, we consider static embeddings to be adequate for our purpose.

Using these embeddings, we calculated pairwise cosine similarities for our list of mathematical terms with all other words in the corpus. Where A and B are the embedded representations of two words, cosine similarity is calculated as

$$\cos(A,B)=AB||A||\ ||B||$$

Intuitively, embeddings of related words - words that appear in similar contexts - have higher cosine similarity. We would expect mathematical vocabulary missing from our initial list to have high cosine similarities with at least one term in our list. When we ranked words by their maximum cosine similarity with a math term, we did not find any missing mathematical terms among the top-ranked words. We note that this method only identifies unigrams (single words), not multi-word expressions or phrases.

Though our primary measures are simple sums of utterance-level counts, we explored the sensitivity of our results to different measurement strategies. To do so, we constructed normalized measures, using type-token ratios, which are the number of mathematical terms used divided by the total number of words spoken in the lesson. Additionally, we constructed an H-index based measure. None of our results differed substantively across these measures, so we report results throughout based on the simple measure for ease of interpretability.

Here is a complete list of the mathematical vocabulary terms used to construct the measures in the paper. Note that, as described in the methodology section, words were stemmed to match multiple forms of that word. This also means that words that reduce to the same stem (e.g. “addition” and “addend”) were considered the same vocabulary term. For some terms, an initialism is sometimes used (e.g. “GCD” for “greatest common denominator”); these are presented separately in the list below but our analysis considered them as the same term. We present here full words, rather than stems, for readability. Hence, multiple forms of the words below were counted by our algorithm (e.g. “hundredth” and “hundredths” were both counted).

Table S1: Mathematical Vocabulary Terms (in Alphabetical Order)

absolute value	composite	equivalence	interquartile range	number sentence
acute	continuous	equivalent	intersect	numerator
add	conversion factor	equivalent expression	interval	numerical
addend	convert	equivalent fraction	iqr	obtuse
addition	coordinate	estimate	isosceles	octagon
algorithm	cube	even number	kilogram	odd number
analog	cubed	expanded form	kilometer	opposite vertex
angle	cubic units	exponent	kite	ordered pair

angle measure	customary system of measurement	expression	lcm	ounce
approximate	customary unit	factors	least common multiple	parallel
arc	data	fraction	length	parallelogram
area	decimal	fraction form	line	parentheses
area model	decimal divisor	fractional unit	line plot	partial product
array	decimal expanded form	frequency	line segment	partition
associative	decimal fraction	gallon	liter	pentagon
attribute	decompose	gcd	long division	percent
average	degree	geometry	make ten	perimeter
axis	denominator	gram	mass	perpendicular
bar graph	dependent variable	graph	mean absolute deviation	picture graph
base ten	diagonal	greatest common factor	measure	pint
benchmark fraction	digit	half circle	measure of center	place value
bisect	dimension	half-circle	median	polygon
bisector	distance	hashmark	meter	polyhedron
box plot	distribution	height of	meters per second	positive number
capacity	distributive	heptagon	metric	pound
categorical data	divide	hexagon	metric unit	prime number
centimeter	divisible	hierarchy	milliliter	prism
circle	division	histogram	millimeter	product
coefficient	divisor	horizontal	mixed number	protractor
collinear	dot plot	hundred	mixed unit	pyramid
common denominator	double number line diagram	hundredth	multiply	quadrant
common factor	endpoint	identity property	multiple	quadrilateral
common multiple	equal group	independent variable	multiplier	quart
commutative	equation	inequality	number path	quarter circle
ratio	tenth			
rational number	tessellate			
reciprocal	tetromino			
rectang	thermometer			
rectangle	thousand			
regular polygon	thousandth			
remainder	total			

repeated addition
rhombus
right angle
rotate
ruler
scale
scaled graph
scalene
semi-circle
semicircle
simplif
simplify
solid figure
solution
square
square unit
square units
squared
standard form
statistic
straight angle
subtract
sum
supplementary angle
surface area
survey
symmetry
trapezoid
triangle
unit
unit cube
unit form
unit fraction
unit interval
unit price
unit rate
unit square
unknown
value
variability
variable
vertex
vertical
volume
whole
whole number
whole unit
word form
yard

Appendix B: Experimental Balance

In the typical educational field experiment, we can test for balance by comparing the pre-treatment covariates of students in the treatment group to those in the control and conducting t-tests. However, in the case of random assignment to teachers, the typical balance checks are not feasible since there are, in effect, many arms of treatment (i.e. each teacher is a separate treatment condition). Recall that, in the NCTE experiment, students were randomly assigned to teachers within school-grades. In this case, experimental balance would imply that, within each school grade, students of different teachers have approximately the same distribution of each pre-treatment covariate we observe.

To test this, we fit the following regression model – separately, for each randomization block – for each pre-treatment covariate.

$$Y_{ig} = \beta_0 + \lambda_g + \varepsilon_{ig}$$

Here, Y_{ig} is the pre-treatment covariate for student i in the class of teacher g. These pre-treatment covariates include prior math test scores, special education status, race/ethnicity indicators, gender, free and reduced priced lunch status, and English language learner status. λ_g is a vector of fixed effects for the teachers in the randomization block. The F-test for these regressions tests whether the teacher fixed effects are jointly predictive of the pre-treatment covariate. Under a balanced experiment, we expect to reject the F-test (at a .05 level) 95% of the time. Because we are conducting these tests for many blocks, we expect some rejections even under the null. To account for this, we adjust the resulting p-values using a Benjamini-Hochberg procedure to control the false discovery rate at 5%.

While students appear to be balanced across teachers on most pre-treatment characteristics, we find evidence of imbalance on prior math test scores in 13 of the 42 blocks. This indicates that, in some school-grades, at least some students were sorted into classrooms non-randomly.

Inclusion of prior math test scores in our regression model of effects on student test scores attenuated our results (these attenuated results are reported in the main paper). This indicates that other researchers using the experimental year of the NCTE data to estimate causal effects should be sure to include lagged test scores in their regression models.