Inside the Black Box: Using Machine Learning to Predict and Understand Effective Teaching

Ben Workman, *MIT* *

August 10, 2026

Abstract

This paper applies machine learning techniques to classroom transcripts of mathematics instruction to predict and understand teacher value-added. Using data from approximately 300 U.S. elementary school teachers, I show that transcript embeddings predict value-added better than observable teacher characteristics, including classroom observation scores, explaining an order of magnitude more out-of-sample variation. I present evidence suggesting these predictions could improve personnel decisions and student outcomes. I then identify specific instructional features associated with value-added, such as the use of student teamwork. Finally, I examine human evaluators' predictions of value-added and find suggestive evidence of errors in their weighting of instructional practices.

*I am grateful to Lawrence Katz for his guidance and support. I thank Raj Chetty, Grace Michel, Sendhil Mullainathan, Ashesh Rambachan, Suproteem Sarkar, Jesse Shapiro, Andrei Shleifer, and participants at workshops at Harvard and MIT for helpful comments. I thank Heather Hill, Mark Chin, and David Blazar for assistance with the NCTE Main Study data. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2141064. E-mail: benworkm@mit.edu.

1 Introduction

A substantial body of evidence has demonstrated that individual teachers have meaningful, quantifiable effects on K–12 student outcomes (e.g., Rockoff 2004; Rivkin et al. 2005; Kane and Staiger 2008; Chetty et al. 2014a; Chetty et al. 2014b; Jackson 2018). Given the importance of individual teachers, much effort has been made to identify characteristics of teachers and their teaching that predict gains in student achievement, both to help schools make important personnel decisions (i.e., whom to hire and to whom to grant tenure) and to inform teacher training curricula (Hanushek 1971; Hanushek et al. 2005; Hill et al. 2005; Kane et al. 2008; Boyd et al. 2009; Staiger and Rockoff 2010; Harris and Sass 2011; Araujo et al. 2016; Bau and Das 2020). Yet across the multitude of studies aiming to do this, researchers have consistently been able to explain only a relatively small amount of the variation in teacher value-added (Hanushek 2011; Jackson et al. 2014; Lovenheim and Turner 2017).[1] Many of these studies include as predictors of teacher quality not only basic information like teacher education and experience but also variables like teacher IQ (e.g., Araujo et al. 2016), personality traits (e.g., Rockoff et al. 2011), and content knowledge (e.g., Bacher-Hicks et al. 2018).

Given the inability of this wide range of variables to explain much of the variation in teacher value-added, several large-scale projects were launched in the early 2010s to gather videos of teachers' instructional practices and score these classroom observations according to rubrics created by education researchers (e.g., the Classroom Assessment Scoring System developed by La Paro and Pianta (2003)). In principle, these observations should reveal what makes some teachers more effective in the classroom than others. In practice, however, measures of classroom instruction derived from video observations are often less correlated with teacher value-added than are other predictors, leading some researchers to conclude that "the search for predictors of teacher effects may proceed more fruitfully along other pathways" (Bacher-Hicks et al. 2018).

Though the human-scored classroom observations have shown limited ability to predict teacher value-added, the raw classroom data may nonetheless contain information relevant to understanding the determinants of effective teaching. Human evaluations of the classroom data can only capture the dimensions of effective teaching already known by researchers, but there may be additional aspects of effective teaching not incorporated

[1] Hanushek notes that "literally hundreds of research studies have focused on the importance of teachers for student achievement," yet "it has not been possible to identify any specific characteristics of teachers that are reliably related to student outcomes." Jackson et al. argue that "observable characteristics are unlikely to be able to predict most of the variation in teacher effects." Lovenheim and Turner state that "we know very little about how to find high-value-added teachers using their observed characteristics."

in existing theories. Supervised machine learning techniques have proven useful in identifying patterns overlooked by human expertise and theories across a wide variety of settings in the social and natural sciences, from judge bail decision prediction (Ludwig and Mullainathan 2024) to early breast cancer detection (McKinney et al. 2020) to earthquake aftershock forecasting (DeVries et al. 2018). Applying these techniques to classroom data could thus reveal important aspects of effective teaching that are missing from human evaluations.

In this paper, I investigate this possibility using data from the National Center for Teacher Effectiveness (NCTE) Main Study, which studied 317 teachers in four large urban school districts in the U.S. from 2010 to 2013 (Kane et al. 2015). I observe a rich set of information about each teacher, including their educational history, teaching experience, content knowledge, and classroom observation scores from mathematics lessons. In addition to the classroom observation scores, I also have the raw transcripts from the classroom observation videos (Demszky and Hill 2023). Finally, I have estimates of teachers' value-added on students' state math test scores, computed by conditioning on prior-year test scores and student and peer characteristics.

I represent each classroom transcript in a low-dimensional embedding space using a text embedding algorithm from OpenAI. This algorithm can represent complex features of teachers' instruction that may matter for student learning but that may not be well represented by a simpler representation, such as the output of a bag-of-words model. I then fit a penalized linear model to learn the relationship between the transcript embeddings and teacher value-added on state math test scores. I separately fit a model to learn the relationship between the wide set of traditional "structured" variables, which includes the human-generated classroom observation scores, and teachers' value-added.

The out-of-sample predictions formed from the transcript embeddings explain over 10 percent of the variation in shrunken value-added estimates and over 5 percent of the variation in unshrunken estimates out of sample, an order of magnitude more than the predictions formed from the structured data. To put these magnitudes in context, the transcript-based predictions explain roughly one-third as much variation in unshrunken value-added as does a prior-year value-added estimate itself. This predictive relationship persists across a variety of alternative modeling choices. Moreover, in a smaller sample of teachers randomly assigned to classroom rosters in the final year of the NCTE Main Study, the estimated relationships between predictions formed using pre-random-assignment transcripts and student achievement are positive, although the estimates are imprecise. This provides suggestive evidence that the predictions capture persistent differences in teacher effectiveness rather than merely nonrandom student sorting.

The value-added predictions derived from the transcript embeddings may be useful in their own right, as the personnel decisions school districts face require making predictions about teacher quality and are thus examples of "prediction policy problems" where improved predictive ability could yield considerable benefits (Kleinberg et al. 2015; Chalfin et al. 2016).[2] I consider two policy-relevant applications in which more accurate value-added predictions could be useful. The first is teacher hiring. Improving the quality of teachers hired has the potential to improve student outcomes substantially, in part because effective hiring decisions allow schools to avoid the costs of ineffective teachers (Staiger and Rockoff 2010). To help inform these consequential decisions, many districts include a mock lesson as part of their hiring process, often with real students.[3] I ask whether applying my value-added prediction method to this demo lesson could meaningfully improve the quality of teachers hired. As a proof of concept, I simulate hiring from the sample of teachers in the NCTE Main Study and compare two hiring policies: one policy that selects teachers based on value-added predictions formed using all "structured" information observable at the time of hire, including human evaluations of one lesson, teacher experience, and a measure of teacher content knowledge, and a second policy that selects teachers based on predicted value-added formed using the transcript embedding from the same single lesson. I show that, relative to the first policy, using the transcript embeddings for teacher selection could indeed improve the quality of teachers hired. For example, if a school district aimed to hire the top 50% of teachers, selecting this set of teachers using the second policy would increase mean value-added among the hired teachers by approximately 0.25 standard deviations relative to the first policy. A back-of-the-envelope calculation using the estimates of the economic value of higher teacher value-added from Chetty et al. (2014b) implies that this improvement would increase cumulative lifetime earnings by approximately $9,692 per student taught by those teachers for one school year. Importantly, this approach assumes that the information in demo lessons is as predictive of teacher value-added as is the information in lessons taught by incumbent teachers. Although I am able to offer some suggestive evidence in support of this assumption—showing that the results are similar across novice and more experienced teachers and across fall and spring lessons—it is not directly testable, and so these results are best viewed as a proof of concept that invites further study.

A second policy-relevant use of these value-added predictions is in predicting future effectiveness after observing a teacher in a classroom for a small number of years, which

[2] Indeed, Kleinberg et al. (2015) cite teacher value-added prediction as an "illustrative example" of the "prediction policy problems" they write about.

[3] Jacob et al. (2018) review the literature on teacher hiring processes and report that the fraction of districts using sample lessons ranges from 15% to 50% depending on the particular survey.

may be useful for informing teacher tenure decisions (Goldhaber and Hansen 2010; Chalfin et al. 2016). Past work, including Cantrell and Kane (2013) and Araujo et al. (2016), has generally found that classroom observation scores are not significant predictors of future value-added once one observes an estimate of past value-added. I show that applying my algorithmic approach, rather than the traditional human approach, to evaluating teachers' performance in the classroom changes this: predictions formed from the transcript embeddings from the first two years of the NCTE Main Study predict student achievement and value-added in the final year, even controlling for estimated value-added in the first two years. These results suggest that the value-added predictions formed from the transcript embeddings could play a useful role in teacher retention decisions—though, naturally, their value decreases as one observes value-added estimates from a larger number of years.

Beyond pure prediction, one may wish to understand what features of the classroom transcripts represented by the black-box embedding algorithm are most related to teachers' value-added. To answer this question, I use sparse autoencoders (Le et al. 2012; Cunningham et al. 2023; Bricken et al. 2023), which decompose embeddings into sparse, interpretable features and have been proposed as a tool for hypothesis generation (Movva et al. 2025; Carlson 2026). I train SAEs to learn a dictionary of features useful for reconstructing the transcript embeddings and aggregate each feature's activations to the teacher level. To identify the features most related to teacher value-added, I use integrated path stability selection (Melikechi and Miller 2025), a technique for selecting features in high-dimensional data while controlling false discoveries, and use an open-source large language model to generate natural-language descriptions of the selected features. This exercise identifies features *correlated* with value-added that may warrant further study, rather than establishing that these features causally affect student learning.

Three features stand out. The first, which describes "sitting or chairs," typically appears when teachers are directing students' behavior (e.g., "sit up," "have a seat") and is negatively related to value-added, consistent with it serving as a symptom of student misbehavior and weak behavior management. The second, which describes "the organization of students into specific teams or groups," is positively related to value-added and corresponds to the use of structured team activities in the classroom, often in the context of team-based academic competitions. The third, which describes situations in which a "teacher explicitly invites a student to come up to the front of the classroom or to the board," is negatively associated with value-added. These relationships hold even after controlling for structured variables, including classroom observation scores, selected by double lasso (Belloni et al. 2013). Overall, the selected features describe aspects of

teachers' classroom management rather than mathematical content, and the teamwork and student presentation features suggest that practices that distribute active work across many students simultaneously are associated with student learning. These practices are not entirely new concepts: existing research and classroom observation instruments recognize the importance of classroom management, student interactions, and eliciting student answers. The value of the approach is instead to highlight more fine-grained practices not captured by existing observation instruments, such as structured team-based activities or particular forms of answer elicitation.

In the final part of this paper, I explore why the information in the classroom transcripts that is predictive of teacher value-added is not better represented in standard classroom observation instruments. The answer does not appear to be that effective teaching depends on too many features to encode in an observation rubric: relatively sparse models fit on the SAE features outperform models fit on all classroom observation variables. I instead focus on the possibility that the designers of the observation instruments place too much weight on some dimensions of teaching relative to their importance for student learning. I study this possibility using two sources of suggestive evidence. First, I take advantage of the fact that education researchers ranked a subset of the NCTE Main Study teachers based on the researchers' predictions about the teachers' value-added on math test scores (Hill et al. 2013). I show that, when forming predictions about teachers' value-added, the researchers substantially overweight the importance of the "richness" of the mathematics instruction relative to its relationship with teacher effects on student test scores. Second, I show that, when forming holistic assessments of the quality of mathematics instruction in a given lesson, raters in the NCTE Main Study place substantial emphasis on the richness of mathematics despite the fact that it is only weakly related to value-added. This empirical evidence, combined with qualitative evidence about the experts' decision-making process, is consistent with the model of human judgment in Bordalo et al. (forthcoming): educational experts may categorize the task of predicting teacher effectiveness as a problem of identifying rich mathematics instruction, leading them to attend to mathematical richness rather than factors related to classroom management and student engagement.

This paper makes two primary contributions. I contribute to the literature on teacher quality by demonstrating that teacher value-added is substantially more predictable than past work has suggested (e.g., Hanushek 2011). My results imply that teachers' behaviors and language use inside the classroom contain meaningful signal about their effectiveness at raising student test scores, validating the use of classroom observations even if the way in which these observations are traditionally conducted does not fully leverage

the information available. Additionally, my discovery of interpretable features that help explain variation in teacher value-added provides new insights into the education production function and suggests specific channels through which teachers may affect student learning. Finally, whereas past work like Staiger and Rockoff (2010) has found little evidence that information available at the time of hire can help school districts improve their personnel selection, my results suggest a possible way to meaningfully improve the quality of teachers hired.[4] Likewise, my results suggest a way to use classroom observations to improve the quality of predictions of teachers' future value-added, which has generally not been possible with human evaluations of teachers' performance in the classroom (Cantrell and Kane 2013; Araujo et al. 2016).

This paper also contributes to a nascent literature on algorithmic hypothesis generation in the social sciences (Ludwig and Mullainathan 2024; Mullainathan and Rambachan 2024; Batista and Ross 2024; Movva et al. 2025; Liang forthcoming). I offer one of the first applications of algorithmic hypothesis generation in the social sciences and, to the best of my knowledge, the first application of these techniques to the question of teacher effectiveness.[5] Understanding the determinants of effective teaching is of great academic and policy interest, and there has been much work examining the correlates of teacher effectiveness (Hanushek 2011). Thus, I have a large set of human-generated hypotheses against which I can benchmark any potential algorithmic discoveries.

The remainder of this paper is organized as follows. Section 2 describes the data I use from the NCTE Main Study. Section 3 outlines my approach to predicting teacher value-added using the classroom transcripts and evaluates the quality of these predictions relative to predictions formed using the existing "structured" data. Section 4 explores two potential policy-relevant applications of improved value-added predictions: teacher hiring and future-value-added forecasting. Section 5 identifies interpretable correlates of value-added used by the black-box predictor. Section 6 examines human judgments of teacher effectiveness. Section 7 concludes.

[4] One of the "five facts about teacher effectiveness" presented by Staiger and Rockoff (2010) is that "school leaders have very little ability to select effective teachers during the initial hiring process."

[5] There has been some recent work in the education literature that attempts to use natural language processing (NLP) techniques to automate the scoring of classroom transcripts and then correlate these algorithmically-generated ratings with teacher value-added metrics (Liu and Cohen 2021; Demszky and Hill 2023). My project takes a fundamentally different approach. Rather than starting from theories about what effective teaching looks like and attempting to operationalize them using NLP, I take the raw text data as input and test whether the text itself predicts teacher value-added.

2 Data

2.1 Background & Institutional Setting

The National Center for Teacher Effectiveness (NCTE) Main Study (Kane et al. 2015) was a partnership between researchers and four large East Coast urban school districts that aimed to study the relationship between instructional practices and student outcomes in elementary school. The study was conducted over the 2010–2011, 2011–2012, and 2012–2013 school years and included 317 teachers and over 10,000 students. Blazar (2024) documents that the teachers and students participating in the study appear similar in terms of demographics to non-participating teachers and students. During the first two years of the study, students were matched to teachers through the districts' usual processes; in the final year of the study, researchers randomly assigned 66 teachers to classroom rosters.[6]

Researchers collected student demographic data and scores on standardized tests in math and English. Using these data, they constructed value-added estimates based on state math and English test scores as described in Section 2.2. In addition, researchers collected a wide range of variables related to teacher background and knowledge, as described in Section 2.3. Finally, teachers participating in the study submitted three videos per year of their classroom instruction in mathematics.[7] Trained raters then evaluated the lesson videos according to two rubrics: the Classroom Assessment Scoring System (CLASS) developed by La Paro and Pianta (2003) and the Mathematical Quality of Instruction (MQI) rubric developed by Hill et al. (2008). Transcripts derived from these classroom videos are made accessible to researchers by Demszky and Hill (2023). Section 2.4 offers more detail on these transcripts.

2.2 Teacher Value-Added Data

2.2.1 Construction of Value-Added Estimates

Kane et al. (2015) provide estimates of teachers' causal effects on state mathematics exam scores computed using up to five years of student test score data (the 2008–2009 school

[6] In the random-assignment analysis sample studied by Bacher-Hicks et al. (2019), 132 of the 316 teachers who participated in at least one year of the study were present in the final year. Among these teachers, about 40 percent either were not interested in participating in random assignment or taught in schools whose leadership was not interested in participating in random assignment. Additionally, a small fraction of teachers were not eligible for inclusion in the random assignment component of the study for various reasons (e.g., they did not teach a class where at least five students had baseline standardized test scores).

[7] Teachers chose which lessons to film over the course of the year. The videos teachers submitted had no impact on how schools evaluated or rewarded them. The average length of submitted videos is approximately one hour.

year through the 2012–2013 school year) from students in schools with teachers who participated in the NCTE Main Study.[8] They estimate the model

$$a_{i,j,k,t} = A_{i,t-1}^{\top}\alpha + S_{i,t}^{\top}\beta + P_{j,k,t}^{\top}\delta + E_{i,t}^{\top}\rho + \mu_k + \theta_{k,t} + \epsilon_{i,j,k,t}, \tag{1}$$

where $a_{i,j,k,t}$ represents the rank-based mathematics state test score for student i in class j taught by teacher k during school year t; $A_{i,t-1}$ captures prior student performance; $S_{i,t}$ includes student demographics; $P_{j,k,t}$ represents peer influences; $E_{i,t}$ contains district and grade-year fixed effects; μ_k and $\theta_{k,t}$ capture teacher and teacher-year random effects, respectively, and $\epsilon_{i,j,k,t}$ is an idiosyncratic error term.[9] Kane et al. (2015) estimate this model jointly on all years of data using Hierarchical Linear Modeling with a teacher random effect μ_k and a teacher-year random effect $\theta_{k,t}$ nested within teacher. The resulting estimate $\hat{\mu}_k$ of μ_k is a shrunken estimate of teacher k's value-added. Kane et al. (2015) also estimate the model separately on individual years to construct year-specific value-added estimates.[10]

My analysis in this paper also makes use of unshrunken estimates of teachers' value-added, which are not provided in the NCTE Main Study dataset. To compute these unshrunken estimates, I estimate the model

$$a_{i,j,k,t} = A_{i,t-1}^{\top}\alpha + S_{i,t}^{\top}\beta + P_{j,k,t}^{\top}\delta + E_{i,t}^{\top}\rho + \mu_k + \epsilon_{i,j,k,t} \tag{2}$$

with μ_k as a teacher fixed effect.[11] I also compute year-specific unshrunken estimates by estimating Equation (2) separately on each school year's test score data.[12] Ideally, I would compute these unshrunken estimates on the same sample used by Kane et al. (2015) to compute the shrunken estimates so that the two sets of estimates are directly comparable. However, the student test score data made available to researchers cover only the 2010–2011, 2011–2012, and 2012–2013 school years and only include students whose teachers participated in the study. In practice, however, shrunken estimates computed using the data available to me align closely with those provided by Kane et al. (2015).[13] This suggests, though does not guarantee, that the restricted sample produces estimates similar to what would be obtained on the full sample.

[8] This includes data from students whose teachers did not participate in the NCTE Main Study.

[9] Please consult Appendix Section D.1.1 for more details on the specific variables in this model.

[10] In this case, the teacher-year effect $\theta_{k,t}$ is omitted from the model.

[11] I remove district fixed effects from $E_{i,t}$ to prevent collinearity with teacher fixed effects.

[12] I omit $P_{j,k,t}$ from the model because these class and cohort characteristics are collinear with the teacher fixed effects: each teacher teaches a single class in a given school year.

[13] Please see Appendix Section D.1.2 for more information.

2.2.2 Validity of Value-Added Estimates

A key question is whether these value-added estimates are unbiased. Intuitively, if the estimates are biased by student sorting on characteristics unobserved by the econometrician, then attempts to predict μ_k may inadvertently discover predictors of the component of $\hat{\mu}_k$ that reflects the source of bias.

Following Chetty et al. (2014a), I consider two relevant notions of bias. An estimator of teacher value-added is *forecast unbiased* if students randomly assigned to a teacher with estimated value-added one unit above that of another teacher score one unit higher on average. A value-added estimator is *teacher-level unbiased* if it converges to that teacher's true value-added as estimation error goes to zero. Unless otherwise stated, my analysis assumes teacher-level unbiasedness. Teacher-level unbiasedness is difficult to assess empirically, and almost all of the literature to date has focused on assessing the degree of forecast bias in value-added models. It is possible to have value-added estimates that are forecast unbiased but biased at the teacher level, but, as Chetty et al. (2014a) note, such a case is knife-edge.[14] I thus briefly present evidence on forecast bias in the NCTE Main Study data and other similar settings while acknowledging that I cannot directly measure the degree of teacher-level bias.

Blazar (2018) and Bacher-Hicks et al. (2019) estimate value-added models similar to those used here using the NCTE data and find that the resulting value-added estimates are forecast unbiased.[15] Their finding is consistent with prior work showing that conditioning on prior-year test scores and student demographics is sufficient to generate forecast-unbiased value-added estimates (e.g., Kane and Staiger 2008; Cantrell and Kane 2013; Chetty et al. 2014a).[16] Following a similar approach, I test for forecast bias in the NCTE-provided shrunken estimates using the random assignment of students to teachers in the 2012–2013 school year, as described in Appendix Section D.1.3. The results, visualized in Appendix Figures A1 and A2, provide no evidence of forecast bias.

2.3 Teacher Characteristics Data

Kane et al. (2015) provide detailed information about the teachers in the sample, including

[14] In particular, the covariance of true value-added and teacher-level bias must exactly equal the negative of the variance of teacher-level bias.

[15] Bacher-Hicks et al. (2019) account for noncompliance by using the effectiveness of a student's randomly assigned teacher as an instrument for the effectiveness of that student's actual teacher. Blazar (2018) alternatively restricts the analysis to classrooms with little to no noncompliance and similarly finds no evidence of forecast bias.

[16] See Rothstein (2017) and the reply by Chetty et al. (2017) for additional discussion of what one can conclude from Chetty et al. (2014a) about bias in value-added models.

their years of experience, educational backgrounds, demographics, and mathematical content knowledge. Additionally, I observe the human-generated ratings from the classroom observations on the many dimensions defined by the CLASS (La Paro and Pianta 2003) and MQI (Hill et al. 2008) instruments. The former instrument focuses on evaluating overall classroom interactions and pedagogical practices, irrespective of the subject matter, while the latter targets specific competencies in mathematics instruction. To construct classroom observation variables at the teacher level, I first average raters' scores within each lesson and then average these lesson means across all of a teacher's lessons. I also compute each teacher's mean years of experience for the years in which they appear in the data.

Because some variables are missing for a sizable fraction of teachers, I focus on the set of teacher characteristics with fewer than 5% missing values among the set of teachers with value-added scores, and I additionally include teachers' mean years of experience. There are 59 such characteristics, which are listed in Appendix Section D.2. Appendix Table D2 presents summary statistics for a subset of these variables on the sample of 278 teachers for whom none of these variables are missing.[17]

I also generate indices corresponding to broad categories of items measured during the classroom observations.[18] These categories are Emotional Support, Instructional Support, Classroom Organization, Student Engagement, Math Richness, Working with Students, Errors and Imprecision, and Student Participation in Meaning-Making and Reasoning. The first four indices are derived from the CLASS items, while the last four indices are derived from the MQI items. Appendix Table D3 identifies the individual items used to construct each index. For ease of interpretation, I take the negative of the "Errors and Imprecision" index and call this "Mathematical Precision."

2.4 Teacher Transcript Data

Demszky and Hill (2023) provide transcript data for a total of 1,660 mathematics lessons from 317 distinct teachers. For the vast majority (approximately 90%) of teachers, I observe at least three transcripts, and for some teachers, I observe as many as 11 tran-

[17] Rather than reporting summary statistics for each classroom observation category, I instead report summary statistics for the whole-lesson MQI score, which reflects raters' assessment of the overall quality of the lesson.

[18] To form each index, I standardize the component variables at the lesson level, take the first principal component of the corresponding variables, orient it so that its average loading on the underlying items is positive, standardize the resulting lesson-level principal component scores, and then average these scores across a teacher's lessons.

scripts.[19] Appendix Figure D1 presents the first 30 rows of the transcript data as prepared by Demszky and Hill (2023). To convert these data from their original "utterance"-level format into one long piece of text per lesson, I iterate through "utterances" and concatenate the speaker label, a colon, and the utterance itself.[20]

3 Predicting Teacher Value-Added

I now turn to predicting teachers' value-added on students' state math test scores, first using the classroom transcripts (Section 3.1) and then using the traditional "structured" data (Section 3.2). Let y_k^{EB} denote teacher k's shrunken (empirical Bayes) estimate of math value-added and y_k^{FE} denote teacher k's unshrunken (fixed effect) estimate of math value-added.

3.1 Predicting Value-Added from Classroom Transcripts

I use OpenAI's text-embedding-3-large embedding algorithm to form numerical representations of the classroom transcripts that I then use to predict value-added on state math test scores.[21] Because many classroom transcripts exceed the embedding model's 8,192 token limit, I split each transcript into sequential segments of up to 8,192 tokens (without splitting utterances) and represent each resulting segment s as a vector $X_s \in \mathbb{R}^{3072}$ using the embedding algorithm.

For each teacher k, I use the embeddings from all teachers except k as vectors of covariates in a ridge regression model. For a transcript from teacher $\ell \neq k$ in year t, the outcome being predicted is the inverse-variance-weighted average of teacher ℓ's year-specific shrunken value-added estimates from all years other than t, $y_{\ell,-t}^{EB}$.[22] I use the shrunken value-added estimates as labels rather than the unshrunken estimates both because they

[19] Following the suggestion of Demszky and Hill (2023), I exclude from the sample two lessons whose transcripts have inaccurate speaker attribution.

[20] For example, the first four utterances in Appendix Figure D1 are concatenated to form "teacher: Friends, yesterday we started off by working on some word problems together, yes? student: Yes. teacher: And yesterday towards the end of the period, you got a word problem given to you that you had to draw a picture to show it, and you had to do the math problem to show it, yes? multiple students: Yes."

[21] This embedding model is pre-trained with a contrastive objective, where the goal is to minimize the distance in embedding space between pieces of text taken from the same source while maximizing the distance in embedding space between pieces of text taken from different sources.

[22] As is common practice when estimating ridge regression models, for each leave-one-teacher-out split, I standardize the embedding dimensions to have mean 0 and variance 1 using the training segments and apply the same transformation to the held-out teacher's segments.

are less noisy and because I observe them for a larger number of years.[23] I choose to predict $y^{EB}_{\ell,-t}$ rather than y^{EB}_{ℓ} (computed using all available years) or $y^{EB}_{\ell,t}$ (computed using only the transcript year) to more directly isolate the stable features of instruction predictive of value-added. When a teacher is unusually effective in a given year, transcripts may capture elevated student engagement or teacher enthusiasm that reflect conditions specific to that year rather than the teacher's persistent instructional quality. Because teacher effectiveness is correlated over time, focusing on $y^{EB}_{\ell,-t}$ still provides a meaningful measure of the teacher's general ability without conflating it with these transient, year-specific patterns.

I then use the model estimated on the transcripts from all teachers except k to predict value-added for all transcript segments from teacher k and average these predictions at the teacher level, weighting by the number of words in the corresponding transcript segment.

Formally, let S denote the set of all transcript segments and $S_k \subset S$ denote those from teacher k. For each segment $s \in S$, let $\ell(s)$ and $t(s)$ denote the corresponding teacher and year, respectively. Then the prediction of teacher k's value-added formed from the "unstructured" transcript data, which I denote as $\hat{y}^u_k$, is given by

$$\hat{y}^u_k = c^u_k + \frac{\sum_{s \in S_k} w_s X_s^\top \theta_k}{\sum_{s \in S_k} w_s}$$

where (c^u_k, θ_k) solve

$$\arg\min_{c,\theta} \left\{ \sum_{s \in S \setminus S_k} \left(y^{EB}_{\ell(s),-t(s)} - c - X_s^\top \theta \right)^2 + \lambda_k \|\theta\|_2^2 \right\}$$

and λ_k is chosen with leave-one-segment-out cross-validation over the training segments in $S \setminus S_k$ and a mean squared loss function.[24]

This procedure yields predictions for all teachers who have at least one transcript that can be linked to a leave-year-out value-added measure. The procedure also ensures that the model parameters used to form a prediction for any given teacher were not estimated using data from that teacher, thereby allowing me to evaluate the out-of-sample performance of my model—which is especially important in this setting given the large number

[23] As noted in Section 2.2, the NCTE provides up to five years of shrunken value-added estimates, whereas the student test score data I use to construct the unshrunken estimates only covers three years. These additional years of data are especially useful given that I do not use the value-added estimate from year t.

[24] I use leave-one-segment-out rather than teacher-level cross-validation to take advantage of the efficient leave-one-out cross-validation implementation in scikit-learn's `RidgeCV`. The reported out-of-sample performance is still evaluated on held-out teachers.

of predictors relative to the number of observations. I repeat this exercise with linear regression (i.e., imposing $\lambda_k = 0$ for all k).

3.2 Predicting Value-Added from Structured Data

To predict value-added using the "structured" information for teacher k—i.e., the 59 teacher characteristics described in Subsection 2.3—I estimate a lasso model using the structured data from all teachers except k and use the parameters of this model to make a prediction for teacher k.[25] Formally, let Z_k denote the vector containing the "structured" information for teacher k and let y_ℓ^{EB} represent teacher ℓ's shrunken (empirical Bayes) math value-added estimate computed using all available years. Then the prediction $\hat{y}_k^s$ is given by

$$\hat{y}_k^s = c_k^s + Z_k^\top \eta_k$$

where (c_k^s, η_k) solve

$$\arg\min_{c,\eta} \left\{ \sum_{\ell \neq k} \left(y_\ell^{EB} - c - Z_\ell^\top \eta \right)^2 + \lambda_k \|\eta\|_1 \right\}$$

and λ_k is chosen by cross-validation on the training teachers in each leave-one-teacher-out split. I repeat this exercise with linear regression, ridge regression, and gradient-boosted trees to ensure that my results are not overly sensitive to the model class used for prediction.[26]

3.3 Evaluating the Predictions

For every teacher k, I now have two predictions of their value-added: $\hat{y}_k^u$, the prediction formed using the "unstructured" data (i.e., the transcripts), and $\hat{y}_k^s$, the prediction formed using the "structured" data. To evaluate whether the transcript data adds explanatory power above and beyond the structured data, I residualize y_k (either the shrunken measure y_k^{EB} or the unshrunken measure y_k^{FE}, depending on the specification), $\hat{y}_k^s$, and $\hat{y}_k^u$ on district and grade fixed effects and estimate the following three regressions on the set of

[25] As is common practice when estimating lasso models, for each leave-one-teacher-out split, I standardize the predictors to have mean 0 and variance 1 using the training teachers and apply the same transformation to the held-out teacher.

[26] When using ridge regression for this prediction exercise, I likewise standardize the predictors to have a mean of 0 and a variance of 1. For ridge regression and gradient-boosted trees, I choose hyperparameters via cross-validation on the training teachers in each leave-one-teacher-out split. I implement the gradient-boosted tree model in Python using scikit-learn's histogram-based gradient boosting class `HistGradientBoostingRegressor`.

teachers for whom both $\hat{y}_k^u$ and $\hat{y}_k^s$ are non-missing:

$$\tilde{y}_k = \alpha + \psi_s \widetilde{\hat{y}}_k^s + \epsilon_k \tag{3}$$

$$\tilde{y}_k = \alpha + \psi_u \widetilde{\hat{y}}_k^u + \epsilon_k \tag{4}$$

$$\tilde{y}_k = \alpha + \psi_s \widetilde{\hat{y}}_k^s + \psi_u \widetilde{\hat{y}}_k^u + \epsilon_k. \tag{5}$$

Table I presents the results of these regressions, where the first three columns use shrunken value-added and the remaining columns use unshrunken value-added as the dependent variable. We observe that the predicted value-added formed from the transcripts is positively associated with both measures of value-added ($p < 0.01$ for shrunken value-added and $p < 0.01$ for unshrunken value-added) and explains over 10 percent of the variation in the shrunken value-added scores (adjusted $R^2 = 0.108$) and over 5 percent of the variation in the unshrunken estimates (adjusted $R^2 = 0.056$). The predictions formed from the structured information are marginally related to shrunken value-added ($p = 0.10$) but are not statistically significantly related to unshrunken value-added ($p = 0.24$) and explain much less of the variation across teachers (adjusted $R^2 = 0.007$ for shrunken value-added and adjusted $R^2 = 0.003$ for unshrunken value-added). Additionally, once we control for the predictions formed from the transcripts, the predictions formed from the structured data are no longer significant. To put these adjusted R^2 values in context, note that a regression of unshrunken value-added on shrunken value-added from the 2009–2010 school year, after partialling out district and grade fixed effects, has an adjusted R^2 of approximately 0.19.

To visualize the quality of these two sets of predictions, Figure I presents binned scatterplots of teachers' estimated value-added vs. the predictions formed from the structured data and the transcript embeddings. Panels A and B use shrunken value-added as the dependent variable, while Panels C and D use unshrunken value-added. We observe that the points in Panels B and D are more tightly clustered around the regression line relative to those in Panels A and C, respectively, consistent with the results reported in Table I.

3.4 Alternative Predictive Models

One might wonder how the quality of the predictions depends on the modeling choices made. I therefore repeat the prediction exercise using alternative model classes, embedding algorithms, value-added labels, and out-of-sample splits. The central finding—that transcript embeddings predict value-added substantially better than the structured data—is robust across these exercises.

First, the results are not driven by the model classes used in the baseline analysis. Appendix Tables A1 and A2 report results using alternative prediction algorithms to form the structured-data- and embedding-based predictions, respectively. None of the predictions formed from the structured data using linear regression, ridge regression, or gradient-boosted trees are statistically significantly related to either measure of estimated value-added. When I use linear regression to predict value-added from the transcript embeddings, the resulting predictions are still significantly associated with estimated value-added, though these predictions explain much less of the variation across teachers (adjusted R^2 = 0.022 for shrunken value-added and 0.011 for unshrunken value-added).[27]

Second, the results are not specific to OpenAI's embedding algorithm. I repeat the prediction exercise using embeddings from Google's Gemini Embedding 2 model; an off-the-shelf version of Google's EmbeddingGemma model, a much smaller (308M-parameter) open-weight model; and a version of EmbeddingGemma that I fine-tune on the classroom transcripts with the goal of having the embedding space better distinguish teachers from one another.[28] Appendix Table A3 reports the results. The predictions formed from the Gemini embeddings are, if anything, slightly better than those formed using the OpenAI embeddings (adjusted R^2 = 0.132 and 0.064 for shrunken and unshrunken value-added). The predictions formed from the EmbeddingGemma embeddings are still significantly related to both measures of estimated value-added but explain less of the variation across teachers (adjusted R^2 = 0.055 and 0.010 for shrunken and unshrunken value-added, respectively), consistent with the model's smaller size leading to lower-quality representations of the classroom dialogue. However, once I fine-tune EmbeddingGemma, the resulting predictions improve substantially (adjusted R^2 = 0.094 and 0.028 for shrunken and unshrunken value-added, respectively).

Third, the results are relatively robust to using unshrunken rather than shrunken value-added as the label being predicted. As shown in Appendix Table A4, when I train the model to predict unshrunken value-added computed using all years except the transcript year, the resulting predictions are marginally associated with unshrunken value-added (p = 0.10; adjusted R^2 = 0.011). Because the unshrunken value-added estimates are available for only three school years, these labels are formed using at most two years of

[27] Intuitively, because the embeddings are very high-dimensional relative to the number of teachers, unregularized linear regression is prone to overfit the training data, so some form of regularization is useful to learn a model that generalizes well to unseen teachers.

[28] I implement this by fine-tuning with a standard contrastive learning objective, treating different lessons from the same teacher as positive pairs and lessons from other teachers in the same district and grade as negative pairs. I perform this fine-tuning in a leave-one-district-out manner: for each held-out district, I fine-tune a separate copy of the model on transcripts from teachers in the other districts and use that copy to embed transcripts from teachers in the held-out district.

data and are especially noisy. If I instead predict unshrunken value-added computed using all available years of data, the resulting predictions are highly statistically significant ($p < 0.01$) and explain more variation in unshrunken value-added (adjusted $R^2 = 0.086$) than do the predictions reported in Table I. The predictions formed from the structured data, by contrast, are not significant.[29]

Fourth, the predictive relationship generalizes across districts. The baseline approach leaves out one teacher at a time and trains the model on all remaining teachers, including teachers from the same district as the held-out teacher. To assess how well the model generalizes across districts, I repeat the exercise but leave out one full district at a time. To isolate the effect of excluding same-district teachers from the mechanical effect of a smaller training sample, I construct a placebo distribution by repeatedly partitioning the teachers into random groups matching the size distribution of the actual districts and re-running the leave-one-group-out prediction on each random partition. Appendix Figure A4 compares the leave-one-district-out adjusted R^2 to this placebo distribution. For both shrunken (adjusted $R^2 = 0.074$) and unshrunken (adjusted $R^2 = 0.032$) value-added, the true leave-one-district-out value is unremarkable relative to the placebo distribution (at the 32nd and 25th percentiles, respectively), implying that excluding same-district teachers does not meaningfully degrade the quality of predictions beyond what would be expected from the smaller training sample.

The appendix contains the results of several additional analyses. Appendix Table A5 shows that allowing the effect of teacher experience to vary over time by repeating the prediction exercise at the teacher-year level does not improve the predictive power of the structured variables. Appendix Figure A3 documents that increasing the number of transcripts used to form teachers' value-added predictions generally results in increased explanatory power. Appendix Table A6 reveals that teacher speech is substantially more predictive than is student speech (adjusted R^2 of 0.066 vs. adjusted R^2 of 0.014).

A final consideration is whether treating this as a supervised learning problem—i.e., training a model to learn the relationship between transcript embeddings and value-added—is necessary at all. After all, cutting-edge large language models (LLMs) are able to perform a wide variety of complex tasks (OpenAI et al. 2024; Bubeck et al. 2025; Luong et al. 2025; Patwardhan et al. 2025), which has engendered hope that LLMs will be helpful in the kinds of prediction tasks of interest to economists (Ludwig et al. 2026). To assess how well off-the-shelf LLMs can predict teacher quality from the transcripts alone, I ask two LLMs, OpenAI's GPT 5.4 and Google's open-source Gemma 4, to predict a teacher's ef-

[29] These structured-data predictions are formed using linear regression, as lasso shrinks all coefficients to zero and thus yields essentially constant predictions.

fectiveness at raising student test scores, on a scale of 1 to 5, using the transcript text.[30] I aggregate the resulting transcript-level predictions to the teacher level. Appendix Table A7 presents regressions of both measures of value-added on the LLM-generated predictions. Both LLMs' predictions have some degree of predictive power for unshrunken and shrunken value-added. However, for both models and both value-added measures, the amount of variance explained by each LLM's predictions is at most around half the amount explained by the predictions formed from the transcript embeddings.

3.5 Transcript Predictions and Student Achievement Under Random Assignment

One potential concern is that the relationship between the transcript-based predictions and value-added scores reflects the nonrandom sorting of students to teachers rather than differences in true teacher effectiveness. To explore this possibility, I again leverage the random assignment of teachers to student rosters in the 2012–2013 school year. I compare the relationship between the transcript predictions and student achievement in the 2012–2013 school year separately for students who participated in the random assignment exercise and for those who did not. Intuitively, if the transcript predictions largely pick up on characteristics of the students a teacher typically teaches, one would expect them to be substantially less predictive of student achievement when teachers are randomly assigned to student rosters. This analysis is relatively low-powered, however, because only a small number of teachers participated in random assignment.[31]

Specifically, I estimate student-level regressions of standardized 2012–2013 state math test scores on standardized teacher value-added predictions formed from transcripts created during the two preceding school years.[32] I estimate these regressions in an observational sample and a random assignment sample. The former consists of students outside the random assignment exercise: neither their student roster nor their actual teacher participated in the exercise. For these students, I use the prediction associated with each student's actual teacher and include district and grade fixed effects and student demographic

[30] Appendix B.1 presents the exact prompt.

[31] There are 54 teachers who were randomly assigned to classrooms in the final year of the NCTE Main Study, have an available value-added prediction, and are in a randomization block in which at least one other teacher has an available value-added prediction. Blocks containing only one such teacher are excluded because they provide no within-block variation in the predicted value-added variable used as the instrument.

[32] For each held-out teacher, I fit the prediction model using only first- and second-year transcripts from the other teachers and then apply it only to the held-out teacher's first- and second-year transcripts. The label used to train this model is shrunken value-added excluding the final study year.

and lagged-test-score controls. The random assignment sample consists of students on rosters to which teachers were randomly assigned. To account for noncompliance, I estimate the coefficient of interest using 2SLS, in which the prediction associated with each student's actual teacher is instrumented by the prediction associated with their randomly assigned teacher. This specification includes randomization-block fixed effects, student demographic and lagged-test-score controls, and mean baseline characteristics of students on the assigned roster. I repeat this exercise separately for the baseline predictions formed from the OpenAI embeddings, for the predictions formed from the Gemini and contrastively-fine-tuned EmbeddingGemma embeddings, and for "ensemble" predictions formed by averaging these three sets of predictions and then re-standardizing them to have unit variance.

Figure II presents the results. All four random assignment point estimates are positive and, if anything, larger than the corresponding observational estimates. Moreover, the random assignment coefficients for the Gemini and the ensemble predictions are statistically distinguishable from zero ($p = 0.015$ and $p = 0.045$, respectively), whereas those for the OpenAI and EmbeddingGemma predictions are not. The estimates are imprecise, however, and there is substantial overlap between the confidence intervals for the various predictive models. For example, the OpenAI estimate is the smallest of the four, but its confidence interval includes the ensemble estimate, which is the largest at 0.098.

Taken together, these results provide no evidence that the relationship between the transcript-based predictions and student achievement weakens when teachers are randomly assigned to student rosters and instead provide suggestive evidence that the predictions capture persistent differences in teacher effectiveness. Given the small random assignment sample and the resulting imprecision, however, they do not establish that the observational relationship is entirely attributable to causal teacher effects.

In the next section, I consider possible policy uses of the enhanced predictive capacity documented thus far, before turning to the question of interpretability in Section 5.

4 Policy Analysis

One reason that having better value-added predictions is useful is that it enables schools to make better personnel decisions. Given that teachers play a central role in the education production function, increasing teacher quality through better personnel policy has the potential to generate substantial economic gains (Chetty et al. 2014b). I consider two specific decisions schools face: teacher hiring and teacher retention.

4.1 Teacher Hiring

When making teacher hiring decisions, schools typically observe information about prospective teachers' educational histories and experience. Additionally, many school districts conduct "demo lessons" in which teachers demonstrate their teaching ability, often in front of actual students. Moreover, large school districts like the ones in the NCTE Main Study often receive a considerable number of applicants from which to choose.[33]

Even with these sources of information and a large set of prospective teachers from which to choose, identifying effective teachers at the time of hire has proven difficult (Staiger and Rockoff 2010).[34] Since observable teacher characteristics are only weakly related to teacher value-added (TVA), even decision-makers who optimally weigh the various pieces of information have limited ability to identify high-value-added teachers.

One may naturally ask whether applying the value-added prediction method developed in Section 3 to the demo lesson could help school districts identify promising applicants. To test whether this is the case, I simulate hiring from the sample of teachers included in the NCTE Main Study, choosing one lesson per teacher at random to serve as an analogue of a demo lesson. Although this exercise provides a reasonable approach to addressing this question, it is important to consider several limitations. First, the teachers in my sample may differ from those that schools encounter at the time of hire. One reason is that I only observe teachers who were hired, meaning that any findings may not generalize to the broader population of applicants. Another reason is that teachers in my sample are likely more experienced on average than applicants. To partially address this concern, I carry out the exercise separately for inexperienced and experienced teachers. Second, the randomly chosen lessons may not be representative of a true demo lesson scenario. While demo lessons often involve teaching in front of real students, the classroom transcripts in my data come from lessons taught when teachers have an established relationship with students. It is possible that there are aspects of teaching that are associated with student learning but that only reveal themselves once a teacher has spent a sufficient amount of time with a group of students. While this concern cannot be fully resolved with my data, I partially address it by comparing the hiring gains from lessons taught earlier with those from lessons taught later in the school year. Similarly, it is possible that the relationship between teacher practices and student learning changes when teacher practices are observed in a high-stakes setting like a demo lesson. Third, I ignore

[33] For example, Jacob et al. (2018) document that only 13 percent of applicants to D.C. Public Schools were ultimately hired.

[34] Staiger and Rockoff (2010) note that "there is scant evidence that school districts or principals can effectively separate effective and ineffective teachers when they make hiring decisions."

any equilibrium effects of school districts' improved ability to identify high-value-added teachers (e.g., Bates et al. 2024).

With these caveats in mind, I now proceed with the exercise. I randomly choose one lesson per teacher and form predicted value-added using only that lesson.[35] To predict value-added using the structured data, I assume that the school district has all the information about a teacher's background used in Section 3.2, as well as the human-generated classroom observation scores for the randomly chosen lesson. I then form predictions from this modified set of covariates using linear regression.[36]

With these two predictions, I can then consider how the distribution of teacher value-added would change if I used the predictions formed from the transcript embedding rather than the more traditional measures. I first consider what would happen if schools selected the top half of teachers within each district as ranked by predicted value-added. Panel A of Figure III plots the distribution of unshrunken value-added for the resulting set of teachers, with one distribution for selection by the transcript embedding and another for selection by the structured data. We observe that the value-added distribution of teachers selected using the transcript embeddings has more mass on the right side (i.e., the higher value-added side), suggesting an improvement in teacher quality.

To more rigorously quantify the benefit of using the transcript embeddings for teacher selection and to understand how such benefits depend on the fraction of teachers selected, I consider hiring some fraction f within each district for $f \in \{0.10, 0.20, ..., 0.90\}$, comparing selection by the transcript embedding to selection by the structured data. The estimand of interest is the expected difference in mean unshrunken value-added between the two selected sets of teachers, where the expectation is taken over repeated groups of teachers drawn from the district populations represented in the NCTE sample and over the random choice of one observed lesson per teacher to serve as the demo lesson, conditional on the fitted prediction models. I estimate this expectation by averaging the difference in mean unshrunken value-added across 200 independent draws of one lesson per teacher. I compute 95% confidence intervals using a within-district bootstrap with 1,000 replicates, averaging over 200 lesson draws within each replicate and treating the lesson-level predictions as fixed.[37] For this estimand to correspond to the difference in

[35] As in Section 3.1, to form a prediction for teacher k, I use transcripts from all lessons from all teachers except k to estimate the model. The only difference is that rather than forming predictions for each transcript from the held-out teacher and then taking the mean of these predictions to form a teacher-level prediction, I only predict for a single randomly chosen transcript.

[36] I use linear regression rather than lasso for this exercise because lasso selects the null model in nearly all splits, yielding essentially constant predictions.

[37] The confidence intervals therefore reflect uncertainty from sampling teachers but not the uncertainty from re-estimating the predictive models.

mean true value-added between the two sets of teachers, we need the average bias in unshrunken value-added to be the same among teachers selected by the transcript-based rule and teachers selected by the structured-data rule. This condition is implied by, but weaker than, teacher-level unbiasedness.

Panel B of Figure III visualizes the results of this exercise. We observe that all nine point estimates are positive, and eight of the nine confidence intervals—for $f = 0.20$ through $f = 0.90$—exclude zero.[38] Relative to teacher selection based on the wide range of variables available to school districts at the time of hire, including human evaluations of a teacher's performance in a lesson, using the embedding predictions can increase the quality of the hired teachers as measured by their effectiveness at raising student test scores. To contextualize the magnitude of these effects, consider that the standard deviation of the value-added variable in the set of teachers used for this analysis is approximately 0.178.[39] Thus the increase in mean value-added for $f = 0.50$ represents an increase of about 0.25 standard deviations in value-added.

To partially address the concern that the teachers in my sample are more experienced than the typical applicant, I repeat this exercise separately for teachers with two or fewer years of experience and for teachers with more than two years of experience. Panel A of Appendix Figure A5 visualizes the results. Though the confidence intervals for inexperienced teachers are relatively wide due to the smaller sample size, all point estimates are larger than those for the experienced teachers (on average across f values, expected gains are 0.053 student-level test-score standard deviations larger for inexperienced teachers than for experienced teachers). Thus, we do not observe evidence that the algorithm is less useful for selecting among inexperienced teachers than for selecting among experienced teachers. Similarly, to partially address the concern that the gains from using the algorithm may be larger when the lesson analyzed is one in which the teacher has an established relationship with students, Panel B of Appendix Figure A5 visualizes the results of repeating this exercise separately for fall and spring lessons.[40] The point estimates are slightly larger for fall lessons than for spring lessons (on average across f values, expected gains are 0.022 student-level test-score standard deviations for fall lessons versus 0.019 for spring lessons), and they generally remain positive. Finally, Appendix Figure A6

[38] Intuitively, as f increases, the number of observations used to compute the increase in value-added increases, increasing the precision of the estimates, yet the policy becomes less effective because we select teachers with lower predicted value-added (and in the extreme case where $f = 1$, the policy has zero effect by definition). This may explain why we do not observe an effect significantly different from zero for the smallest value of f.

[39] I compute this standard deviation by taking the square root of the estimated variance of the teacher random effect μ_k from the value-added model described in Section 2.2.

[40] I restrict this analysis to the set of teachers for whom I observe both fall and spring lessons.

replicates Figure III but uses unshrunken value-added in years other than the demo lesson year as the outcome variable, while Appendix Figure A7 uses only lessons from the first NCTE Main Study year and uses unshrunken value-added in the last two years as the outcome variable. The results are somewhat noisier but broadly similar: for example, for $f = 0.50$, the respective increases in mean value-added are 0.34 and 0.24 in standard deviations of teacher value-added, compared to 0.25 when using all lessons and unshrunken value-added computed using all years.

Is this improvement in teacher quality economically significant? One way to answer this question is by drawing on the estimates made by Chetty et al. (2014b) of the effect of having a higher value-added teacher on students' earnings in adulthood. Chetty et al. (2014b) use data from a large urban school district similar to those studied by the NCTE Main Study, so it is reasonable to assume that their findings generalize to my setting. The authors estimate that a one-standard-deviation increase in value-added for one school year raises each student's cumulative lifetime earnings by about $39,000 in 2010 dollars. Applying this estimate, the simulated 0.25 standard-deviation improvement in the mean value-added of teachers hired under the transcript-based policy implies a gain of approximately $9,692 in cumulative lifetime earnings per student taught by those teachers in a given school year. For a class of 20 students, this corresponds to about $194,000 in cumulative lifetime earnings per classroom-year. These exact magnitudes should be interpreted cautiously, given that they come from a simulated hiring exercise and use estimates of the economic value of teacher value-added calculated in a different setting. Still, they suggest that even modest improvements in predictive accuracy could have economically meaningful benefits. Such benefits would be especially compelling if the marginal cost of applying the algorithm were low, as may be the case in districts already conducting demo lessons. The possibility for economically meaningful net benefits from this predictive algorithm is consistent with recent work by Ludwig et al. (2024) demonstrating that predictive algorithms in fields, including education, can have infinitely large Marginal Values of Public Funds (Hendren and Sprung-Keyser 2020).

4.2 Predicting Future Effectiveness

In addition to teacher hiring, a second key personnel decision that school districts face relates to teacher tenure. Many school districts make tenure decisions after a teacher has been teaching for just a few years. These decisions are consequential: granting tenure to an ineffective teacher could mean decades of decreased student performance. Thus, being able to forecast future effectiveness given only a few years' worth of data is useful.

Though the TVA predictions derived from the transcript embeddings are strongly related to teachers' estimated value-added, they need not be a useful input to this decision: when making this decision, schools not only have access to teacher characteristics like years of experience but also have data on past performance (e.g., past value-added estimates). Thus the TVA predictions derived from the transcripts are only useful inputs to this decision-making process if they provide information above and beyond what is captured in observed value-added estimates.

To test whether transcript predictions provide additional signal, I form two predictions of each teacher's future value-added: one using only transcripts from the first NCTE Main Study year and one using transcripts from the first two years. I take two approaches to forming these predictions. The first is a strict forecast in which, for each held-out teacher and each transcript window, I fit the predictive model using only other teachers' transcripts from the same window. When using transcripts from the first year (2011, i.e., the 2010–2011 school year), the label being predicted is the precision-weighted average of shrunken value-added estimates from 2009, 2010, and 2011; when using first- and second-year transcripts, the label is shrunken value-added excluding the final study year. The second approach is a leave-one-district-out forecast in which, for each held-out teacher, I fit the predictive model on all transcripts from teachers outside the held-out teacher's district, using as the label for each transcript the teacher's shrunken value-added computed excluding that transcript's year. I then apply this model to the held-out teacher's transcripts from the first year or the first two years. While this second approach uses transcripts and value-added labels from future years in other districts, the transcripts to which the estimated model is applied are the same in both approaches.

For both sets of predictions, I then estimate two regressions. First, I regress unshrunken value-added in the final year of the NCTE Main Study (2013) on shrunken value-added in 2011, the prediction derived from the structured data, and the transcript prediction formed using only the teacher's 2011 transcripts. Second, I regress unshrunken value-added in 2013 on shrunken value-added in 2011 and 2012, the prediction derived from the structured data, and the transcript prediction formed using the teacher's 2011 and 2012 transcripts. The goal of these two regressions is to forecast a teacher's value-added in the third NCTE Main Study year using only the information from that teacher that is available in the first year and the first and second years, respectively.

Columns 1 and 2 of Table II report the results of these two regressions for the strict forecasts, while Columns 3 and 4 use the leave-one-district-out predictions. As expected, the shrunken value-added estimates are strongly predictive of unshrunken value-added in 2013. Controlling for these shrunken value-added estimates, the strict transcript-based

forecasts are not associated with unshrunken value-added in 2013. However, the leave-one-district-out prediction formed from transcripts from the first two years is predictive of future unshrunken value-added ($p = 0.07$) after conditioning on shrunken value-added in the first two years. By contrast, the TVA predictions formed from the structured data are insignificant in all specifications, a result consistent with past work in this area (e.g., Cantrell and Kane (2013) show that controlling for one year's value-added estimate renders the coefficients on both student survey scores and classroom observation scores insignificant in a regression predicting out-of-sample value-added for elementary school math teachers).

Columns 5 and 6 provide a student-level analogue of this forecasting exercise. In particular, I estimate Equation (2), in which I model final-year math test scores as a function of student and peer characteristics, but replace the teacher fixed effect with the teacher-level predictors used above: prior shrunken value-added estimates, the predictions formed from structured data, and the strict transcript forecasts. The transcript predictions from both the first year and first two years predict final-year test scores in this model ($p = 0.05$ and $p = 0.01$, respectively).

Thus, while traditional classroom observations have proven to be of little use in forecasting future value-added, the raw classroom data contain useful information predictive of future value-added and student achievement, even controlling for several estimates of past value-added. Insofar as schools aim to grant tenure to teachers in part based on their future effects on student achievement, these predictions could serve as a useful input to schools' highly consequential tenure decisions.

5 Discovering Interpretable Correlates of Value-Added

The policy applications considered in the previous section are settings where prediction is useful in its own right. Beyond pure prediction, one might hope to understand what human-interpretable features of the transcripts are associated with value-added.

As an initial benchmark, I examine whether the transcript-based predictions can be understood through the existing structured variables. To do so, I regress the transcript predictions on the 59 structured variables used to form the structured-data predictions in Table I. Appendix Figure A8 presents the regression coefficients and associated standard errors from this model, which has a leave-one-teacher-out R^2 of 0.075.[41] This limited fit,

[41] After applying the Benjamini–Hochberg procedure to control the false discovery rate at 5%, six coefficients remain statistically significant: Asian, Hispanic, and Black teachers tend to receive higher predictions relative to White teachers, while those with higher "negative climate" scores, those who use a more

together with the substantially greater predictive power of the transcript-based predictions, suggests that the existing structured variables provide only a limited account of the relevant information contained in the transcripts. I therefore turn to a more direct, data-driven approach that constructs interpretable features from the transcript embeddings themselves. In Section 5.1, I provide an overview of my methodology for constructing these features, which I then apply to my dataset in Section 5.2. In Section 5.3, I identify a subset of these features systematically related to value-added. Finally, in Section 5.4, I discuss these findings.

5.1 Methodological Overview

One can think of a black-box embedding model as representing a given piece of text as a linear combination of some set of interpretable feature vectors, an idea known as the linear representation hypothesis (Park et al. 2024).[42] Conceptually, a feature could be as specific as an individual word or a more abstract concept like a focus on student discipline. My goal is to identify a subset of these features whose presence is related to teacher value-added.

To do so, I begin by constructing a *dictionary* of features that appear in the classroom transcripts. Rather than specifying ex-ante which features to include in the dictionary, I use the classroom transcripts to learn a set of features useful for reconstructing embeddings of the classroom speech. I do so using a technique called sparse autoencoders (SAEs), which has successfully been used for unsupervised feature generation in a variety of settings (e.g., Le et al. 2012; Sun et al. 2016; Praveen et al. 2018) and has gained particular attention over the past several years as a promising approach to interpreting the inner workings of large language models (e.g., Cunningham et al. 2023; Bricken et al. 2023; Templeton et al. 2024). Intuitively, given a set of input embeddings and a pre-specified dictionary size (e.g., a dictionary of 512 features), SAEs find a set of feature vectors that best allows one to reconstruct the input embeddings as linear combinations of the feature vectors, subject to the constraint that only a small number of features can be used

"active format," and those who use more "linking and connections" between different ways of representing a given mathematical idea tend to receive lower predicted value-added. The predictions are not simply picking up on teacher demographics, though: in Appendix Table A8, which residualizes the transcript predictions and outcomes on race and gender dummies in addition to school district and grade fixed effects, we observe that the transcript predictions retain most of their predictive power.

[42] A famous example of this phenomenon comes from Mikolov et al. (2013). Using the embedding algorithm from the RNN toolkit (Mikolov 2012), they show that subtracting the embedding for "man" from the embedding for "king" and adding the embedding for "woman" yields a vector very close in embedding space to "queen."

for reconstruction on any given embedding.[43] Concretely, this is done by minimizing the average *reconstruction loss*—the distance between an input embedding and its sparse reconstruction—via gradient descent.

My implementation differs from the basic SAE algorithm in two key ways. First, rather than choosing one fixed dictionary size, I choose a set $\mathcal{M}$ of different sizes and for each $m \in \mathcal{M}$ separately add to the loss function the reconstruction error achieved using only the first m features in the dictionary. This approach, termed "Matryoshka SAEs" (Nabeshima 2024; Bussmann et al. 2025; Zaigrajew et al. 2025) in reference to the nested dictionary sizes, allows one to generate features at varying levels of granularity: because earlier features in the dictionary need to be able to reconstruct the inputs by themselves, these features tend to be high level, whereas later features are more granular. Second, following recent work (e.g., O'Neill et al. 2024), I use an auxiliary loss that discourages features from becoming inactive during training. For a more detailed description of the SAE architecture, loss function, and training procedure, please see Appendix C.

5.2 Feature Generation

To apply the method outlined in Section 5.1 to the classroom transcripts, I begin by splitting the transcripts into short segments, formed by accumulating utterances until reaching approximately 200 tokens, subject to a 50-token minimum. These short segments allow me to more easily isolate individual features than if I were working with the transcripts from entire hour-long lessons. I then represent each segment as a vector $X \in \mathbb{R}^{3072}$ using OpenAI's text-embedding-3-large embedding algorithm. Because SAE training is unsupervised and does not use value-added outcomes, I fit the SAE dictionaries once using all transcript segment embeddings.

To reduce dependence on any single SAE training run, I train ten Matryoshka SAEs using different random seeds with the HypotheSAEs package (Movva et al. 2025). Each SAE outputs a sparse vector of feature *activations* for each segment, where an activation of zero means that a feature was not used in the reconstruction and larger magnitudes indicate a stronger contribution. Because individual SAE features can vary across random seeds, I cluster similar features across runs using both similarity between the learned feature vectors and correlation between their activations across transcript segments. For each resulting cluster, I average the activations of its member features within each segment and then across a teacher's segments to construct teacher-level feature measures. I also

[43] This sparsity constraint is motivated by the intuition that sparsity can achieve efficient and biologically plausible representations (Olshausen and Field 1996).

record whether any member feature in a cluster is active and refer to each cluster by the feature number assigned by the clustering algorithm. Appendix C describes the training, clustering, aggregation, and feature-prevalence procedures in detail.

After constructing this dictionary of SAE features, I test whether the features collectively have predictive power for value-added. To do so, I aggregate the clustered SAE activations to the teacher level and use these teacher-level mean activations as covariates in a lasso model predicting shrunken value-added. As in Section 3.1, I form predictions by holding out one teacher at a time, estimating the model using all other teachers, and then predicting value-added for the held-out teacher. Table III reports the relationship between these predictions and teachers' estimated value-added. The predictions formed from the SAE activations are significantly associated with both value-added measures and, evaluated on the same sample, explain about half as much variation as the predictions fit on the full embeddings analyzed in Table I (adjusted $R^2 = 0.052$ for shrunken value-added and 0.029 for unshrunken value-added). Thus, although the SAE representation sacrifices some predictive information relative to the higher-dimensional embeddings, it retains substantial signal, making it a useful basis for the feature selection exercise in the next section.

5.3 Discovering Predictive Features

To identify the features most related to teacher value-added, I use integrated path stability selection (IPSS; Melikechi and Miller 2025), a technique for selecting features in high-dimensional data while controlling false discoveries. IPSS builds on stability selection (Meinshausen and Bühlmann 2010; Shah and Samworth 2012), which takes a base variable selection algorithm, such as the lasso, applies it repeatedly to subsamples of the data, and selects variables that are chosen consistently across these subsamples. Rather than evaluating stability at a single level of sparsity, IPSS uses information from the full regularization path (i.e., the sequence of lasso models obtained as one varies the penalty parameter). Specifically, IPSS integrates selection frequencies over this path, looking for features that are selected consistently across many subsamples and across a range of values of the penalty parameter. IPSS summarizes this stability in a feature-level q-value that approximately controls the false discovery rate.

In my implementation, I use lasso as the base selector, shrunken value-added as the outcome variable, and focus on features with q-values below 0.20, of which there are five. Appendix C.4 describes the construction of the q-values, the assumptions underlying the false-discovery guarantees, and the remaining implementation details.

At this point, these five features are only unlabeled directions in embedding space. To interpret them, I follow the standard approach in the SAE mechanistic interpretability literature (e.g., Paulo et al. 2025) and use an LLM to generate natural-language descriptions of the selected cluster features. For each cluster, I provide highly activating and non-activating transcript segments to the open-source LLM Gemma 4 and ask it to identify the most specific attribute present in the former but absent from the latter. I then evaluate each description on a separate set of activating and non-activating segments by asking the LLM whether the described feature is present in each segment. I summarize agreement between these ratings and the SAE cluster presence indicator using the F1 score. Appendix C.5 provides the sampling procedure, prompts, and scoring details.

Table IV presents the selected features, the generated feature descriptions, and the corresponding F1 scores, their correlations with both value-added measures, and their IPSS q-values. Appendix Table C1 reports the five highest-activating examples.

Before discussing these features in the next section, I make two observations. First, the IPSS procedure selects features that have predictive power out of sample. To see this, I hold out one teacher at a time, run IPSS on the remaining teachers using teacher-level SAE activations as features and shrunken teacher value-added as the outcome, fit a linear model on the selected features, and predict value-added for the held-out teacher.[44] Appendix Table A9 reports the relationship between these predictions and teachers' estimated value-added. The predictions are significantly associated with and explain variation in both value-added measures (adjusted R^2 = 0.040 and 0.025 for shrunken and unshrunken, respectively). Second, Appendix Table A10 documents that all selected features except feature 1016 (whose description has an F1 score of 0.198) are strongly related to the algorithmic transcript predictions analyzed in Section 3.1, indicating that the selected features help explain what aspects of the transcripts the predictive model from Section 3 uses.

5.4 Analyzing Discovered Features

I now turn to discussing the discovered transcript features presented in Table IV. Because these relationships are correlational, I interpret the selected features and their relationship to value-added as potential hypotheses warranting further study rather than as evidence that these practices causally affect value-added.

Feature 73, which describes "sitting or chairs," is negatively related to both shrunken

[44] For computational feasibility, for this prediction exercise, I use 250 penalty values and 500 complementary half-samples.

and unshrunken value-added, even after partialling out structured control variables selected by double selection with lasso (Belloni et al. 2013). As is apparent in the examples presented in Appendix Table C1, this feature typically occurs when the teacher is directing students' behavior (e.g., "sit up," "sit properly," "have a seat") and thus may be interpreted as a symptom of student misbehavior and weak behavior management. The connection between behavior management and student learning is not a new insight; indeed, behavior management is a dimension of the CLASS instrument. Still, the feature contains substantial signal beyond the existing CLASS measure and thus could be useful in practice when identifying classrooms with low student learning.

Feature 850, which describes "the organization of students into specific teams or groups," is positively related to both measures of value-added. The general use of teams and group work connects to a literature in education on cooperative learning (Slavin 1980; Johnson and Johnson 1994) and may be a natural channel through which peer effects (Hoxby 2000; Sacerdote 2011) manifest themselves. The closest analogues in the existing classroom observation variables are measures of time spent in small groups versus in an active format, but these measures are only weakly correlated with Feature 850 and do not distinguish between different types of group- or team-based work. Examining the top-activating examples in Appendix Table C1 provides two potential hypotheses about why the types of team-based activities associated with this feature are correlated with value-added above and beyond general group work. First, the examples showcase extremely structured team-based work, with the teachers numbering teams and sometimes assigning specific roles or tasks ("One of you does need to be watching the timer. The other one could be writing out expressions"). Second, many of the top-activating examples in the classroom transcripts involve competitions between teams, such as quiz games in which teams take turns supplying answers, which may improve student learning by using competition as a non-monetary incentive for effort (Levitt et al. 2016).

Feature 680, which describes situations in which a "teacher explicitly invites a student to come up to the front of the classroom or to the board," is negatively related to value-added, even after partialling out double-lasso controls. This relationship may be somewhat surprising given existing work in education arguing that well-executed student board presentations are a key component of high-quality instruction (Stein et al. 2008). There are at least two possible explanations for this finding. First, student board presentations may be effective in the setting I study only when they are carefully planned—for example, when teachers think through which students will present, what each student will present, and in what order—and the teachers in the lessons I study may not execute the presentations in this way. Indeed, in most of the examples shown in Appendix Ta-

ble C1, teachers ask for volunteers to come up to the board rather than having a predetermined plan. Second, because student presentations necessarily leave most students watching rather than actively participating, they may be ineffective in certain settings, such as those where students have difficulty staying focused while their classmate is at the board. We see evidence in the fifth example that student behavior can be an issue during these presentations, when the teacher tells students to "be respectful, please put the caps on the dry erase markers, place the dry erase markers in front of you. Have body control. Your eyes are on the presenter."

The remaining two features warrant less discussion. Feature 854, which describes "the instruction to switch" and is positively correlated with shrunken value-added, could reflect effective management of transitions between activities in the classroom. However, one should avoid reading too much into this feature, as its correlation with unshrunken value-added is essentially zero. Finally, as discussed in Section 5.3, the description of feature 1016 cannot reliably distinguish activating from non-activating transcript segments (F1 score of 0.198), so I make no attempt to interpret it.

Stepping back, several patterns emerge across the selected features. All of the interpretable features describe aspects of teachers' classroom management rather than mathematical content or math-specific pedagogical strategies. Additionally, the signs on the teamwork and student presentation features point in a consistent direction: the student teamwork feature, which is positively related to value-added, reflects a practice that distributes work across many students simultaneously, while the student presentation feature, which is negatively associated with value-added, captures a form of participation that is more concentrated on one student at a time.

Finally, I reiterate that the practices identified here are not wholly new instructional concepts. Existing research and classroom observation instruments recognize that student interactions, elicitation of student answers, and classroom management influence student learning. The potential value of the approach used here is to highlight more fine-grained practices that are not captured in existing classroom observation instruments. For example, rather than asking whether interaction between students is important in general, one can form specific hypotheses about the types of such interactions (e.g., in the form of structured team-based activity) that may be especially important for student learning. Likewise, one can move beyond assessing the overall frequency with which a teacher asks for student answers and instead form hypotheses about whether and how the specific forms of elicitation may matter.

6 Human Judgment

The results presented in Section 3 demonstrate that classroom teaching and interactions do indeed contain meaningful information about teachers' effects on student test scores, and the results presented in the previous section show that this information is at least in part human-interpretable. Yet human-generated classroom observation scores are not strongly predictive of teacher value-added. Given that these classroom observations are conducted in part to identify teacher and student behaviors associated with student learning (Bill & Melinda Gates Foundation 2010), the weak relationship with value-added scores is perplexing. Why do the classroom observations fail to pick up on the features of classroom interactions associated with student learning?

One simple possibility is that student learning depends on a very large number of instructional features—many more than could realistically be encoded in a classroom observation rubric but can be represented in transcript embeddings. However, the results from the previous section suggest that this explanation is insufficient: when predicting value-added from the SAE features using lasso, the maximum number of SAE features used across the leave-one-teacher-out folds was only 14, fewer than the number of classroom observation variables. An alternative explanation is that the creators of teacher observation protocols hold inaccurate beliefs about the characteristics of teachers that affect student learning. For example, if they overestimate the importance of some aspect of teaching for student learning, that aspect of teaching may be overrepresented in the classroom observation rubrics, and vice versa. Understanding these beliefs may thus shed light on the weak relationship between classroom observation scores and teacher value-added.

I use two complementary strategies to study these beliefs. The first leverages the fact that for a small subset of the teachers in the NCTE Main Study, education researchers ranked the teachers according to the researchers' predictions of value-added (Hill et al. 2013). The second strategy is to examine observers' holistic whole-lesson MQI scores, which reflect their assessment of the overall quality of the mathematics instruction. The first strategy has the benefit of directly studying the object of interest but relies on a small sample, while the second approach uses a larger sample but less directly captures observers' beliefs. Sections 6.1 and 6.2 present evidence from these two approaches, and Section 6.3 briefly discusses the results through the lens of a model of selective attention.

6.1 Human Predictions of Value-Added

Twelve researchers, including members of the team that designed the MQI instrument, participated in the exercise of ranking teachers based on value-added predictions. For three of the four districts in the NCTE Main Study, the researchers selected nine teachers from each district—three from the bottom quintile of the value-added distribution, three from the middle quintile, and three from the top quintile. They viewed at least five classroom videos for each teacher and then collectively ranked the teachers according to the researchers' predictions of value-added on mathematics test scores.

The resulting value-added predictions were relatively inaccurate: for 13 of the 27 teachers (48%), the observers predicted the wrong quintile, and 7 of the 18 (39%) teachers in either the top or bottom quintile of value-added were misclassified into the reverse quintile (i.e., teachers in the bottom quintile were ranked in the top quintile, and vice versa).[45] To understand potential reasons for these errors, I examine the relationships between teachers' instructional features and the expert predictions and compare these relationships to the relationships between instructional features and estimated value-added. To measure high-level instructional practices, I use the indices discussed in Section 2.3. Because these indices are derived from the classroom observation scores, they by definition cannot capture aspects of instruction not measured in the classroom observations that may be underweighted by humans, but they can nevertheless be useful in identifying the characteristics associated with human predictions of value-added.

In the prediction data from Hill et al. (2013), I cannot observe human predictions at the teacher level. Instead, in each district, I observe the mean prediction for the three teachers in each of the first, third, and fifth value-added quintiles.[46] Thus, while I cannot examine the relationship between teacher characteristics and expert predictions of value-added, I can examine the relationship between the average characteristics of teachers in a particular value-added quintile in a given district and the average value-added predictions for these teachers.

Panel A of Figure IV presents the difference between two sets of coefficients: coefficients from regressions of the average predicted value-added rank on each of the indices of instructional practices and coefficients from regressions of estimated average value-added rank on the indices. Higher ranks correspond to higher estimated or predicted

[45] For context, predicting at random would in expectation yield error rates of 67% and 33%, respectively, while ranking based on the algorithmic predictions formed from the transcript embeddings in Section 3 would yield 59% and 11%, respectively. The algorithm was trained on transcripts from the same set of lessons viewed by the human observers but, unlike the observers, had no access to auditory or visual information.

[46] Please consult Appendix D.6 for a complete discussion of these data.

value-added. Because the human experts ranked teachers within districts, these regressions include district fixed effects. Additionally, given the small number of observations, I estimate a separate regression for each instructional feature. Positive values indicate that the instructional feature is more strongly associated with predicted value-added rank than with estimated value-added rank.

We observe that the coefficients across the two regressions differ most substantially for the "Math Richness" instructional feature: while this feature is associated with a higher predicted value-added rank, it is associated (albeit statistically insignificantly) with a lower estimated value-added rank. This difference is significant using both a finite-sample t-test ($p = 0.023$) and a within-district permutation test ($p = 0.005$) and remains significant at the 10% level after applying the Benjamini-Hochberg correction ($q = 0.094$ and $q = 0.037$, respectively). We also see some evidence that this difference in coefficients is positive for "Student Participation in Meaning-Making and Reasoning" and "Emotional Support." By contrast, the smallest difference occurs for the "Classroom Organization" feature, which measures skills like behavior management.

6.2 Holistic Mathematical Quality of Instruction Ratings

In addition to looking at the experts' direct predictions of value-added, it is also instructive to examine the holistic whole-lesson MQI scores given to teachers as part of the classroom observations. Just as I identified features of instruction associated with higher human value-added predictions, I can identify features of instruction associated with higher whole-lesson MQI scores and compare these relationships to the relationships between the subcomponents of MQI and value-added. Unlike in the previous section, divergences between the instructional features emphasized by MQI raters and those most predictive of value-added cannot be labeled as "mistakes," since raters are not directly predicting teachers' value-added. Still, if raters tend to prioritize similar features when judging the overall quality of the mathematics instruction, this would lend credibility to the previous section's analysis, which relied on value-added predictions for a small sample of teachers.

Since the whole-lesson MQI scores are by construction focused on the aspects of a lesson relating to the mathematics, I only include in these analyses the four indices derived from MQI variables ("Math Richness," "Student Participation in Meaning-Making and Reasoning," "Mathematical Precision," and "Working with Students"). Additionally, because I have substantially more data than I did for the analysis of expert predictions, I include all four instructional features in the regressions simultaneously (i.e., regressing whole-lesson MQI scores and value-added estimates on the four MQI instructional features plus

district fixed effects). Moreover, I standardize value-added scores and whole-lesson MQI scores to have mean zero and standard deviation one.

Panel B of Figure IV presents the differences between coefficients from a regression of teachers' whole-lesson MQI scores on the instructional features and the coefficients from a regression of teachers' unshrunken value-added estimates on the same features. As with the human value-added predictions, the difference is largest for "Math Richness," and here it is also significantly positive for "Mathematical Precision" but not significantly different from zero for the other two features.

Two supplementary exercises support this interpretation. First, I repeat the feature selection procedure used in Section 5 to identify SAE features predictive of whole-lesson MQI scores. As shown in Appendix Table C2, many of the interpretable features with F1 scores of at least 0.5 capture teachers' emphasis on mathematical reasoning—including discussion of proof and justification, mathematical methods or procedures, and pattern recognition—rather than teacher-student interactions. For these features, whole-lesson MQI is more strongly associated with the features than is unshrunken value-added. Second, Appendix Figure A9 shows that the pattern in Panel B is similar when value-added is measured using the NCTE assessment designed to capture higher-level mathematics skills rather than the state examination.

Taken together, the evidence from the human predictions of teacher value-added and from the whole-lesson MQI scores is consistent with these experts placing greater weight on the richness of mathematics, relative to its association with value-added, and less weight on factors like classroom organization and student engagement. Interestingly, it is these latter aspects of instruction not pertaining to the math quality itself that the approach in Section 5 uncovered. This is not because the algorithm cannot identify "rich" mathematics—indeed, one can explain approximately 40 percent of the variation in the "Math Richness" feature out of sample using the transcript embeddings—but rather because mathematical richness seems to be largely unrelated to student learning in these data.

6.3 Interpreting Human Judgment

Why do experts making value-added predictions and the raters conducting teacher observations tend to heavily weigh factors like the richness of the mathematics? There are, of course, a wide range of theories that could help explain this pattern, and in this paper I will not be able to fully distinguish between them. Qualitative evidence provided by Hill et al. (2013) on the experts' decision-making processes is nevertheless informative.

Teachers frequently "possessed both a range of instructional features and strengths and weaknesses in their implementation, rendering raters unable to prioritize." There were therefore many features to which raters could attend, and the features they emphasized influenced their final predictions. In one case, for example, one rater focused on a "strong classroom climate" and ranked the teacher in the highest quintile, while another focused on "imprecise teaching of content" and ranked the same teacher in the lowest quintile. Which features received attention may have depended in part on how raters categorize the task: in one district, for example, raters "sought to reach agreement by focusing on the degree and quality of inquiry-based instruction," thereby narrowing the set of features they needed to consider.

This qualitative description of raters' decision-making process is consistent with the model presented by Bordalo et al. (forthcoming), in which decision-makers categorize a problem based on similar problems encountered in the past and the categorization of the problem shapes which features are attended to. Viewing the rater decision-making process in light of this model leads one to consider how raters' past experiences influence the way they represent the problem of predicting value-added. One possible explanation for the tendency to focus on features like math richness is that the past experiences of the experts in Hill et al. (2013) and the classroom observation raters may have led them to conceptualize teaching quality primarily in terms of mathematical and pedagogical sophistication, rather than dimensions such as behavior management and classroom organization. Those with greater experience in behavior management might instead conceptualize the task of identifying teachers who help students learn as one of finding teachers who can keep students engaged and on task, which would suggest that they would attend to different features of the same lesson when making predictions about teacher quality. Such possibilities are beyond the scope of this paper but represent promising areas for future work.

7 Conclusion

This paper reveals that there is substantial untapped potential in raw classroom observation data. I demonstrate that text embeddings of classroom transcripts can explain substantially more of the variation in teachers' estimated effects on student test scores than can a wide range of existing variables, including years of experience, teacher content knowledge, and human evaluations of teachers' classroom instruction. This improved predictive ability could improve teacher personnel decisions and thus generate economically meaningful benefits for students. For example, I show that a hiring policy that uses

value-added predictions formed from the transcript of a single lesson could increase cumulative lifetime earnings by $9,692 per student taught by selected teachers for one school year relative to a policy that uses the standard information on job applicants available to school districts. I then use an algorithmic hypothesis generation technique to uncover specific teaching practices—such as the use of teamwork in the classroom—that predict teachers' estimated effects on student test scores, even controlling for observable teacher characteristics. Finally, I examine human experts' predictions of teacher value-added and find that these experts place too much weight on the importance of the "richness" of mathematics in a lesson for improving student test scores, an observation that, combined with qualitative evidence on the experts' decision-making process, is consistent with the model of reasoning and choice proposed by Bordalo et al. (forthcoming).

These results suggest several promising directions for future work. First, given the relatively narrow scope of the dataset used in this paper, it would be valuable to apply the approach developed here to other datasets, including those with student outcomes beyond short-run aggregate test scores—such as item-level outcomes (Bruhn et al. 2025), longer-run test scores (Gilraine and Pope 2021), non-cognitive skills (Jackson 2018), or involvement in the criminal justice system (Rose et al. 2022). Second, one goal of this paper was to identify novel features of teaching *correlated* with student learning. Future work could investigate whether these relationships are causal, and if so, how to use this information to improve teaching—for example, by providing personalized feedback to teachers (Allen et al. 2011). Third, I presented suggestive evidence in Section 4 that the predictive algorithm developed here could be deployed to improve school districts' personnel decisions. One could test this proposition by evaluating how deployment of an algorithm that predicts value-added—for example, from recordings of teacher demo lessons—affects personnel decisions and student learning. Fourth, future work could improve on the interpretability methods used here, given that my SAE implementation did not recover all of the signal contained in the full transcript embeddings. Finally, the analysis in Section 6 on human predictions of teacher value-added focused on the beliefs of academic education researchers. It would be useful to better understand how different stakeholders (e.g., principals, students, and parents) form predictions about teacher quality and what the consequences of these beliefs are. Doing so could both help identify and address biases in human decision-making and help adjudicate between competing behavioral and cognitive explanations for large human prediction errors in this domain.

Tables

Table I: TVA Prediction with Structured Data and Transcripts

	Shrunken TVA			Unshrunken TVA		
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted TVA, Structured Data	0.609*		−0.235	0.815		−0.210
	(0.367)		(0.364)	(0.691)		(0.692)
Predicted TVA, Transcript Embeddings		1.164***	1.222***		1.432***	1.484***
		(0.223)	(0.247)		(0.411)	(0.444)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	261	261	261	261	261	261
R^2	0.011	0.111	0.112	0.007	0.060	0.060
Adjusted R^2	0.007	0.108	0.106	0.003	0.056	0.053

Note: Columns 1 through 3 report OLS regressions of teachers' shrunken (empirical Bayes) value-added on the predictions formed from the structured data (Column 1), the transcript embeddings (Column 2), and both (Column 3). Columns 4 through 6 report analogous regressions using teachers' unshrunken (fixed effect) value-added as the dependent variable. All variables are residualized on school district and grade fixed effects prior to estimation. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table II: Forecasting Future TVA with Structured Data and Transcripts

	Unshrunken TVA, 2013				Student Math Score, 2013	
	(1)	(2)	(3)	(4)	(5)	(6)
Shrunken TVA, 2011	0.342**	0.158	0.334**	0.132	0.442***	0.252**
	(0.144)	(0.151)	(0.140)	(0.146)	(0.106)	(0.112)
Transcript prediction, 2011	−0.164		1.237		0.946*	
	(0.634)		(1.026)		(0.483)	
Shrunken TVA, 2012		0.511***		0.558***		0.518***
		(0.181)		(0.176)		(0.131)
Transcript prediction, 2011–12		0.958		1.697*		1.579**
		(0.805)		(0.944)		(0.633)
Structured data prediction	2.021	0.983	0.664	0.051	−0.081	−0.699
	(2.229)	(1.954)	(2.434)	(2.233)	(1.274)	(1.366)
Observations	69	69	69	69	1,485	1,485
R^2	0.105	0.227	0.130	0.257	0.668	0.671
Adjusted R^2	0.064	0.179	0.089	0.210	0.657	0.660

Note: This table reports OLS regressions testing whether value-added predictions formed from transcript embeddings can help forecast final-year outcomes after controlling for observed shrunken (empirical Bayes) value-added and predicted value-added formed from structured data. Columns 1 through 4 use unshrunken (fixed effect) value-added in the 2012–2013 school year as the dependent variable. As described in Section 4.2, Columns 1 and 2 use strict transcript-based forecasts of value-added, while Columns 3 and 4 use leave-one-district-out forecasts. In Columns 1 and 3, I regress unshrunken value-added in 2013 on shrunken value-added in 2011, the prediction derived from the structured data, and the transcript embedding prediction formed using the teacher's 2011 transcripts. In Columns 2 and 4, I regress unshrunken value-added in 2013 on shrunken value-added in 2011, shrunken value-added in 2012, the prediction derived from the structured data, and the word-count-weighted average of transcript embedding predictions formed using the teacher's 2011 and 2012 transcripts. These four regressions are estimated only on teachers for whom all regressors used in Columns 1 through 4 are non-missing. All variables are residualized on school district and grade fixed effects prior to estimation. Columns 5 and 6 provide a student-level analogue of this forecasting exercise: I estimate Equation (2), in which I model final-year math test scores as a function of student and peer characteristics, but replace the teacher fixed effect with the teacher-level predictors used in Columns 1 and 2. Robust standard errors are reported for Columns 1 through 4; standard errors in Columns 5 and 6 are clustered by teacher. *p<0.1; **p<0.05; ***p<0.01

Table III: TVA Prediction with Sparse Autoencoder Activations

	Shrunken TVA		Unshrunken TVA	
	Main Sample	Full Sample	Main Sample	Full Sample
	(1)	(2)	(3)	(4)
Predicted value-added	1.564***	1.469***	1.998***	1.910***
	(0.341)	(0.329)	(0.691)	(0.673)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes
Observations	261	294	261	294
R^2	0.056	0.047	0.032	0.027
Adjusted R^2	0.052	0.044	0.029	0.024

Note: This table reports OLS regressions of teacher value-added on out-of-sample predictions formed from teacher-level clustered SAE activations, as described in Section 5.2. Predictions are formed by fitting lasso models to predict shrunken value-added using all teachers except one held-out teacher, then predicting value-added for the held-out teacher. Columns 1 and 2 use shrunken value-added as the dependent variable, while Columns 3 and 4 use unshrunken value-added as the dependent variable. Within each outcome group, the first column restricts to the main prediction sample used in Table I, while the second column uses all teachers for whom both transcripts and value-added data are available. All variables are residualized on school district and grade fixed effects prior to estimation. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table IV: IPSS-Selected SAE Features (No Controls and PDS Correlations)

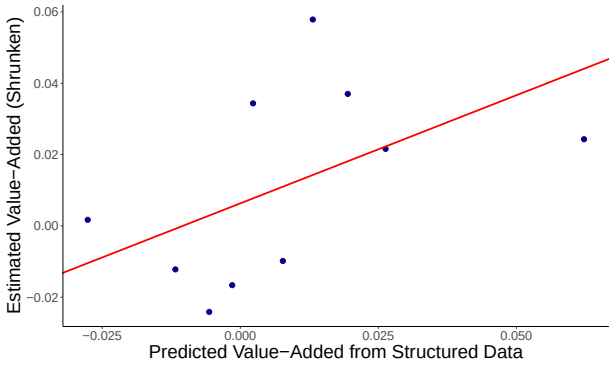
Feature Number	Correlations with Shrunken TVA	Correlations with Unshrunken TVA	IPSS q-value	Description	F1 Score
73	No controls: -0.231 PDS: -0.157	No controls: -0.188 PDS: -0.149	0.047	mentions sitting or chairs	0.990
850	No controls: 0.177 PDS: 0.177	No controls: 0.124 PDS: 0.133	0.047	mentions the organization of students into specific teams or groups	0.936
854	No controls: 0.107 PDS: 0.107	No controls: 0.017 PDS: 0.018	0.094	mentions the instruction to "switch"	0.802
1016	No controls: 0.085 PDS: 0.085	No controls: 0.086 PDS: 0.089	0.192	mentions a mnemonic involving McDonald's and cheeseburgers	0.198
680	No controls: -0.203 PDS: -0.179	No controls: -0.194 PDS: -0.174	0.192	teacher explicitly invites a student to come up to the front of the classroom or to the board	0.980

Note: This table presents the SAE features selected by the IPSS procedure described in Section 5.3. The first column reports the cluster feature number. The second and third columns report correlations between teacher-level mean activations and shrunken and unshrunken teacher value-added, respectively. In these columns, the "No controls" correlation reports the Pearson correlation after residualizing both variables on district and grade fixed effects, while "PDS" (post-double-selection) reports the correlation after also partialling out structured predictors selected by double-selection lasso to predict either the SAE feature or the relevant value-added measure. The fourth column reports the IPSS q-value used for feature selection. The fifth column reports the natural-language description of the feature generated by having the open-source LLM Gemma 4 compare transcript segments where the feature is present to those where the feature is not present. Finally, the sixth column reports the F1 score measuring how well the natural-language description distinguishes activating from non-activating transcript segments, as described in Section 5.3. The score ranges from 0 to 1, with higher values indicating better agreement between the description-based LLM ratings and whether the SAE feature is active in the segment.

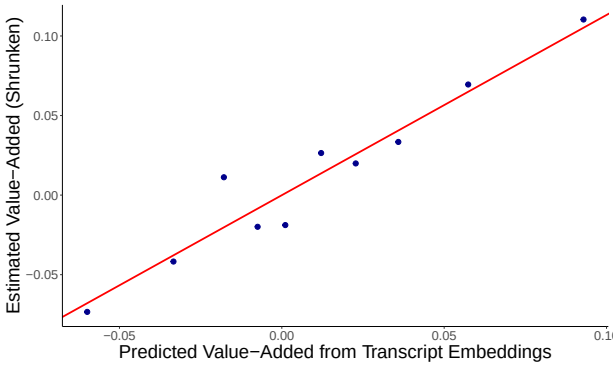
Figures

Figure I: Comparison of Estimated vs. Predicted Teacher Value-Added

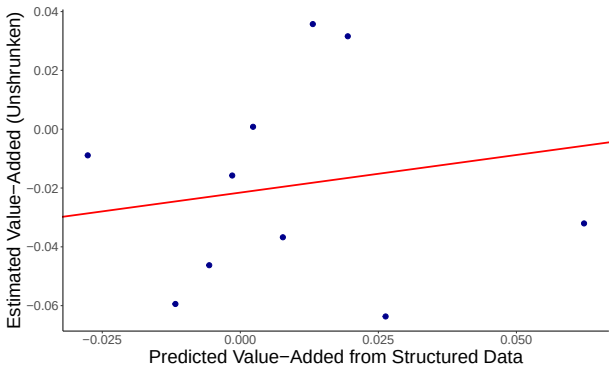
Panel A: Shrunken TVA vs. Predicted TVA from Structured Data



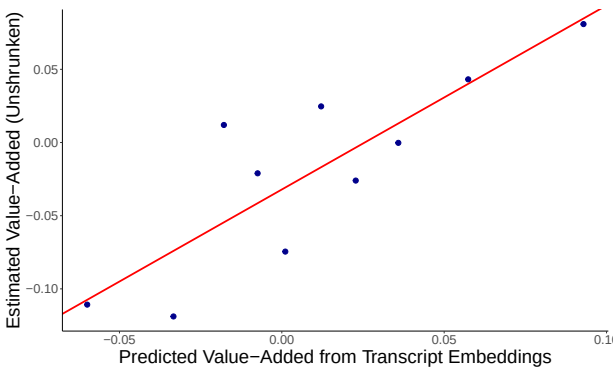
Panel B: Shrunken TVA vs. Predicted TVA from Transcript Embeddings



Panel C: Unshrunken TVA vs. Predicted TVA from Structured Data

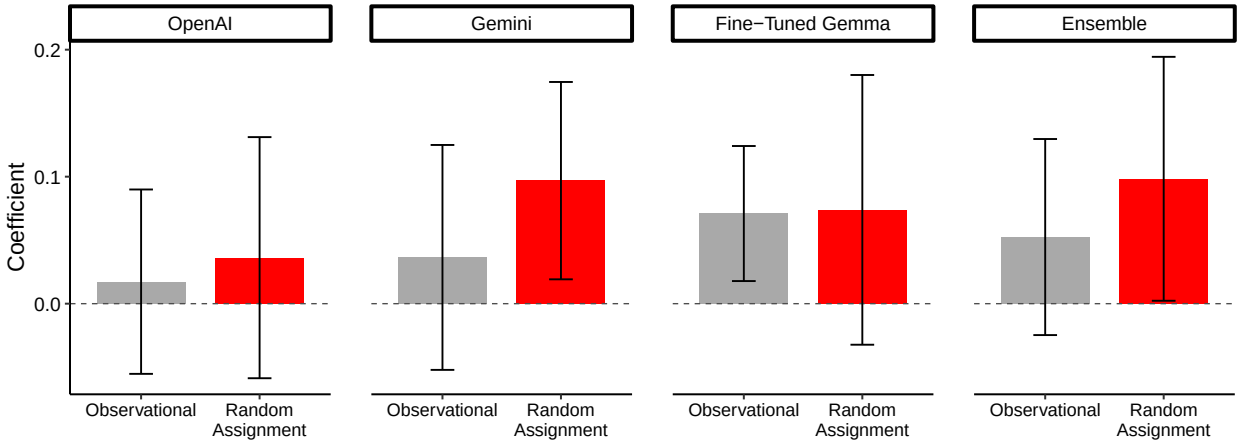


Panel D: Unshrunken TVA vs. Predicted TVA from Transcript Embeddings



Note: This figure presents binned scatter plots of teachers' estimated value-added vs. value-added predictions formed using the structured data and the transcript embeddings. Panels A and B use shrunken value-added as the dependent variable, while Panels C and D use unshrunken (fixed effect) value-added. I split teachers into 10 bins based on predicted value-added and plot the mean value of estimated value-added against the mean value of predicted value-added within each bin. The regression lines are estimated using OLS on the underlying teacher-level data.

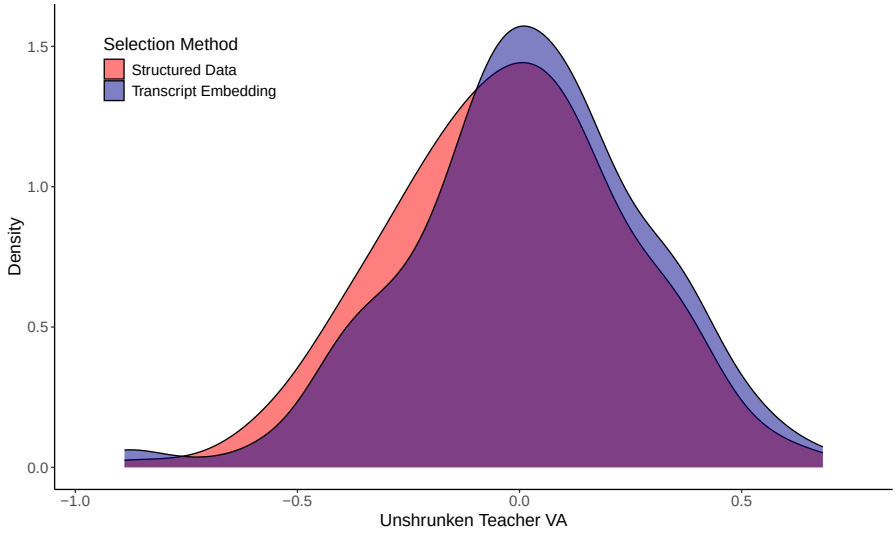
Figure II: Relationship Between Transcript-Based TVA Predictions and Student Achievement Under Random Assignment



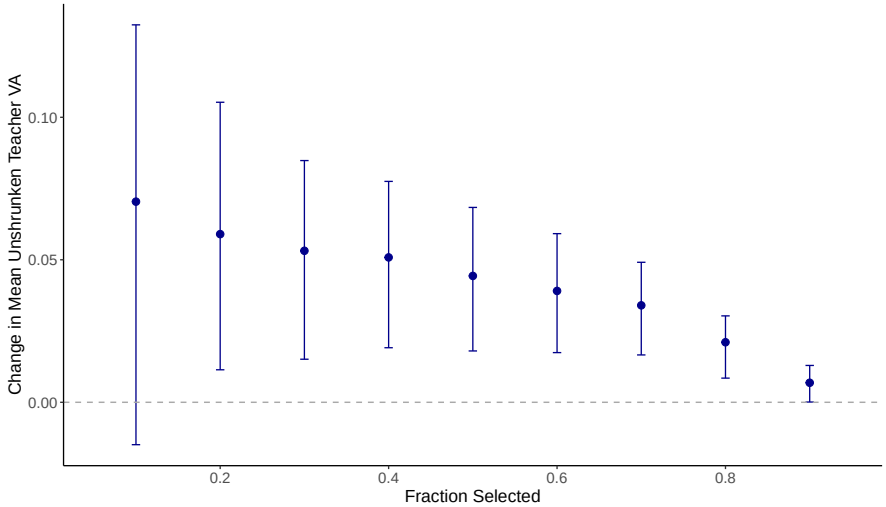
Note: This figure presents estimated coefficients from student-level regressions of standardized state math test scores in the 2012–2013 school year on standardized teacher value-added predictions formed from classroom transcripts created in the two preceding school years. The predictions in the first panel are formed using OpenAI's text-embedding-3-large embedding model, as described in Section 3.1; the predictions in the second and third panels are formed using Google's Gemini Embedding 2 model and a contrastively-fine-tuned version of Google's EmbeddingGemma model, respectively, as described in Section 3.4; and those in the fourth panel are formed by averaging these three sets of predictions and then re-standardizing to have unit variance. The gray bars present estimates for the observational sample, which consists of 2012–2013 students outside the random-assignment exercise: neither their student roster nor their actual teacher participated in the exercise. These regressions use the prediction associated with each student's actual teacher and include district and grade fixed effects and student demographic and lagged-test-score controls. The red bars present two-stage least squares (2SLS) estimates for the random-assignment sample, which consists of students on rosters to which teachers were randomly assigned. To account for noncompliance with the assigned teacher, the prediction associated with each student's actual teacher is instrumented by the prediction associated with the student's randomly assigned teacher. These regressions include randomization block fixed effects, student demographic and lagged-test-score controls, and assigned-teacher roster mean classmate characteristics. Error bars represent 95% confidence intervals. Standard errors are clustered by actual teacher in the observational regressions and by assigned teacher in the random assignment regressions. The dashed horizontal line corresponds to a coefficient of zero. First-stage F-statistics for the OpenAI, Gemini, EmbeddingGemma, and ensemble 2SLS estimates are 754.3, 806.3, 3,902.8, and 2,734.6, respectively.

Figure III: Teacher Hiring

Panel A: Value-Added Distributions Under Two Teacher Selection Policies



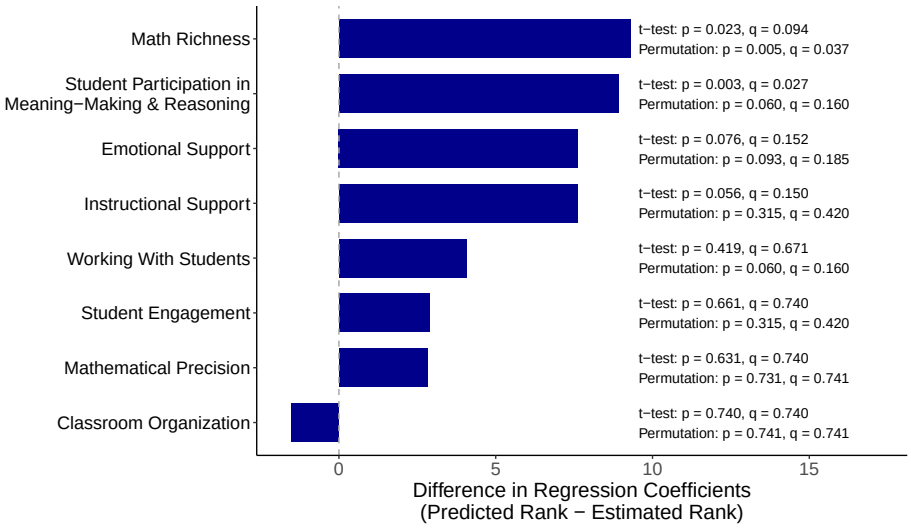
Panel B: Increase in Mean Value-Added Relative to Selection Using Structured Data



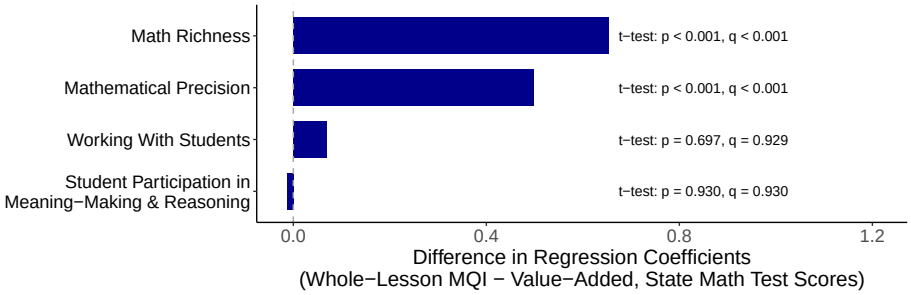
Note: Panel A displays kernel density estimates for two unshrunken value-added distributions: one for teachers in the top half of predicted value-added within each district based on the structured data (red distribution) and another for teachers in the top half within each district based on transcript data (blue distribution). The density estimates are generated using a Gaussian kernel with bandwidth selection following the method provided by Silverman (1998). Panel B presents estimates of the increase in mean unshrunken value-added resulting from using the transcript predictions, rather than the structured data predictions, to choose teachers within each district, for different fractions of teachers hired (as described in Section 4.1). The 95% conditional bootstrap confidence intervals are based on 1,000 within-district bootstrap replicates and 200 lesson draws per replicate, holding the fitted prediction models fixed.

Figure IV: Human Judgment

Panel A: Human Predictions of Value-Added



Panel B: Whole-Lesson MQI Scores and State-Test Value-Added



Note: Panel A plots the difference between the coefficient from a regression of average predicted value-added rank on the instructional feature and the coefficient from a regression of average estimated value-added rank on the same instructional feature. Positive values indicate that the instructional feature is more strongly associated with predicted value-added rank than with estimated value-added rank. These regressions include district fixed effects and are estimated separately by instructional feature for the groups of three teachers used in the value-added prediction exercise described in Section 6. Panel B plots the difference between the coefficient from a regression of teachers' whole-lesson MQI scores on the instructional feature and the coefficient from a regression of teachers' unshrunken value-added estimates on the same instructional feature. Positive values indicate that the instructional feature is more strongly associated with whole-lesson MQI scores than with value-added on state math test scores. The value-added and whole-lesson MQI regressions include all four MQI instructional features together with district fixed effects, as described in Section 6. In Panel A, finite-sample t-test p-values come from regressions of predicted rank minus estimated rank on the instructional feature and district fixed effects; within-district permutation p-values recompute this coefficient after permuting predicted ranks within district. In Panel B, p-values come from regressions of standardized whole-lesson MQI minus standardized value-added on all four MQI instructional features and district fixed effects. The Benjamini-Hochberg q-values adjust these p-values for multiple testing across the instructional features within each panel.

References

Allen, Joseph P., Robert C. Pianta, Anne Gregory, Amori Yee Mikami, and Janetta Lun (2011) "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement," *Science*, 333 (6045), 1034–1037, 10.1126/science.1207998.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady (2016) "Teacher Quality and Learning Outcomes in Kindergarten," *The Quarterly Journal of Economics*, 131 (3), 1415–1453, 10.1093/qje/qjw016.

Bacher-Hicks, Andrew, Mark J. Chin, Heather C. Hill, and Douglas O. Staiger (2018) "Explaining Teacher Effects on Achievement Using Commonly Found Teacher-Level Predictors," *Annenberg Institute EdWorkingPapers*.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger (2019) "An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys," *Economics of Education Review*, 73, 101919, 10.1016/j.econedurev.2019.101919.

Bates, Michael, Michael Dinerstein, Andrew C. Johnston, and Isaac Sorkin (2024) "Teacher Labor Market Policy and The Theory of the Second Best," *The Quarterly Journal of Economics*, 10.1093/qje/qjae042.

Batista, Rafael M. and James Ross (2024) "Words That Work: Using Language to Generate Hypotheses," in *NeurIPS 2024 Workshop on Behavioral Machine Learning*, https://openreview.net/forum?id=pvnXlajGBZ.

Bau, Natalie and Jishnu Das (2020) "Teacher Value Added in a Low-Income Country," *American Economic Journal: Economic Policy*, 12 (1), 62–96, 10.1257/pol.20170243.

Belloni, A., V. Chernozhukov, and C. Hansen (2013) "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81 (2), 608–650, 10.1093/restud/rdt044.

Bill & Melinda Gates Foundation (2010) "The MQI Protocol for Classroom Observations," Technical Report, Measures of Effective Teaching (MET) Project, https://usprogram.gatesfoundation.org/news-and-insights/usp-resource-center/resources/the-mqi-protocol-for-classroom-observations.

Blazar, David (2018) "Validating Teacher Effects on Students' Attitudes and Behaviors: Evidence from Random Assignment of Teachers to Students," *Education Finance and Policy*, 13 (3), 281–309, 10.1162/edfp_a_00251.

——— (2024) "Why Black Teachers Matter," *Educational Researcher*, 53 (8), 450–463, 10.3102/0013189x241261336.

Bordalo, Pedro, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer (forthcoming) "A Cognitive Theory of Reasoning and Choice," *The Quarterly Journal of Economics*.

Boyd, Donald J., Pamela L. Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff (2009) "Teacher Preparation and Student Achievement," *Educational Evaluation and Policy Analysis*, 31 (4), 416–440, 10.3102/0162373709353129.

Bricken, Trenton, Adly Templeton, Joshua Batson et al. (2023) "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning," Transformer Circuits Thread, October, https://transformer-circuits.pub/2023/monosemantic-features.

Bruhn, Jesse, Michael Gilraine, Jens Ludwig, and Sendhil Mullainathan (2025) "Do Test Scores Misrepresent Test Results? An Item-by-Item Analysis," Working Paper 34484, National Bureau of Economic Research, 10.3386/w34484.

Bubeck, Sébastien, Christian Coester, Ronen Eldan et al. (2025) "Early science acceleration experiments with GPT-5," https://arxiv.org/abs/2511.16072.

Bussmann, Bart, Noa Nabeshima, Adam Karvonen, and Neel Nanda (2025) "Learning Multi-Level Features with Matryoshka Sparse Autoencoders," https://arxiv.org/abs/2503.17547.

Cantrell, Steven and Thomas J. Kane (2013) "Ensuring Fair and Reliable Measures of Effective Teaching," Technical Report, Bill & Melinda Gates Foundation, https://usprogram.gatesfoundation.org/-/media/dataimport/resources/pdf/2016/12/met-ensuring-fair-and-reliable-measures-practitioner-brief.pdf.

Carlson, Jacob (2026) "Making Interpretable Discoveries from Unstructured Data: A High-Dimensional Multiple Hypothesis Testing Approach," https://arxiv.org/abs/2511.01680.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016) "Productivity and Selection of Human Capital with Machine Learning," *American Economic Review*, 106 (5), 124–27, 10.1257/aer.p20161029.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a) "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104 (9), 2593–2632, 10.1257/aer.104.9.2593.

——— (2014b) "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104 (9), 2633–79, 10.1257/aer.104.9.2633.

——— (2017) "Measuring the Impacts of Teachers: Reply," *American Economic Review*, 107 (6), 1685–1717, 10.1257/aer.20170108.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey (2023) "Sparse Autoencoders Find Highly Interpretable Features in Language Models," https://arxiv.org/abs/2309.08600.

Demszky, Dorottya and Heather Hill (2023) "The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts," in Kochmar, Ekaterina, Jill Burstein, Andrea Horbach et al. eds. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 528–538, Toronto, Canada: Association for Computational Linguistics, July, 10.18653/v1/2023.bea-1.44.

DeVries, Phoebe M. R., Fernanda Viégas, Martin Wattenberg, and Brendan J. Meade (2018) "Deep learning of aftershock patterns following large earthquakes," *Nature*, 560 (7720), 632–634, 10.1038/s41586-018-0438-y.

Gilraine, Michael and Nolan Pope (2021) "Making Teaching Last: Long-Run Value-Added," Working Paper 29555, National Bureau of Economic Research, 10.3386/w29555.

Goldhaber, Dan and Michael Hansen (2010) "Using Performance on the Job to Inform Teacher Tenure Decisions," *American Economic Review*, 100 (2), 250–255, 10.1257/aer.100.2.250.

Hanushek, Eric A. (1971) "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro-Data," *American Economic Review*, 61 (2), 280–288.

——— (2011) "The economic value of higher teacher quality," *Economics of Education Review*, 30 (3), 466–479, https://doi.org/10.1016/j.econedurev.2010.12.006.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin (2005) "The Market for Teacher Quality," Working Paper 11154, National Bureau of Economic Research, 10.3386/w11154.

Harris, Douglas N. and Tim R. Sass (2011) "Teacher training, teacher quality and student achievement," *Journal of Public Economics*, 95 (7-8), 798–812, 10.1016/j.jpubeco.2010.11.009.

Hendren, Nathaniel and Ben Sprung-Keyser (2020) "A Unified Welfare Analysis of Government Policies*," *The Quarterly Journal of Economics*, 135 (3), 1209–1318, 10.1093/qje/qjaa006.

Hill, Heather C., David Blazar, Andrea Humez et al. (2013) "Examining High and Low Value-Added Mathematics: Can Expert Observers Tell the Difference?" in *Association for Public Policy Analysis & Management Fall Research Conference*, Washington, DC.

Hill, Heather C., Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah L. Ball (2008) "Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study," *Cognition and Instruction*, 26, 430–511, 10.1080/07370000802177235.

Hill, Heather C., Brian Rowan, and Deborah Loewenberg Ball (2005) "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement," *American Educational Research Journal*, 42 (2), 371–406, 10.3102/00028312042002371.

Hoxby, Caroline (2000) "Peer Effects in the Classroom: Learning from Gender and Race Variation," Working Paper 7867, National Bureau of Economic Research, 10.3386/w7867.

Jackson, C. Kirabo (2018) "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes," *Journal of Political Economy*, 126 (5), 2072–2107, 10.1086/699018.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger (2014) "Teacher Effects and Teacher-Related Policies," *Annual Review of Economics*, 6 (Volume 6, 2014), 801–825, https://doi.org/10.1146/annurev-economics-080213-040845.

Jacob, Brian A., Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen (2018) "Teacher applicant hiring and teacher performance: Evidence from DC public schools," *Journal of Public Economics*, 166 (C), 81–97, 10.1016/j.jpubeco.2018.08.011.

Johnson, David W. and Roger T. Johnson (1994) *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning*, Needham Heights, MA: Allyn and Bacon, 4th edition.

Kane, Thomas, Heather Hill, and Douglas Staiger (2015) "National Center for Teacher Effectiveness Main Study," 10.3886/ICPSR36095.V4.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger (2008) "What does certification tell us about teacher effectiveness? Evidence from New York City," *Economics of Education Review*, 27 (6), 615–631, 10.1016/j.econedurev.2007.05.005.

Kane, Thomas J. and Douglas O. Staiger (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," Working Paper 14607, National Bureau of Economic Research, 10.3386/w14607.

Kingma, Diederik P. and Jimmy Ba (2017) "Adam: A Method for Stochastic Optimization," https://arxiv.org/abs/1412.6980.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015) "Prediction Policy Problems," *American Economic Review*, 105 (5), 491–95, 10.1257/aer.p20151023.

La Paro, Karen M. and Robert C. Pianta (2003) "Classroom Assessment Scoring System," 10.1037/t08945-000.

Le, Quoc V., Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng (2012) "Building high-level features using large scale unsupervised learning," https://arxiv.org/abs/1112.6209.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff (2016) "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational

Performance," *American Economic Journal: Economic Policy*, 8 (4), 183–219, 10.1257/pol.20130358.

Liang, Annie (forthcoming) "Using Machine Learning to Generate, Clarify, and Improve Economic Models," *Journal of Economic Literature*.

Liu, Jing and Julie Cohen (2021) "Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods," *Educational Evaluation and Policy Analysis*, 43 (4), 587–614, 10.3102/01623737211009267.

Lovenheim, Michael and Sarah E. Turner (2017) *Economics of Education*: Macmillan Higher Education.

Ludwig, Jens and Sendhil Mullainathan (2024) "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, 139 (2), 751–827, 10.1093/qje/qjad055.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (2024) "The Unreasonable Effectiveness of Algorithms," *AEA Papers and Proceedings*, 114, 623–627, 10.1257/pandp.20241072.

——— (2026) "Large Language Models: An Applied Econometric Framework," *Annual Review of Economics*, 18, 283–316, 10.1146/annurev-economics-120925-105620.

Luong, Thang, Dawsen Hwang, Hoang H. Nguyen et al. (2025) "Towards Robust Mathematical Reasoning," https://arxiv.org/abs/2511.01846.

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole et al. (2020) "International evaluation of an AI system for breast cancer screening," *Nature*, 577 (7788), 89–94, 10.1038/s41586-019-1799-6.

Meinshausen, Nicolai and Peter Bühlmann (2010) "Stability Selection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72 (4), 417–473, 10.1111/j.1467-9868.2010.00740.x.

Melikechi, Omar and Jeffrey W. Miller (2025) "Integrated Path Stability Selection," *Journal of the American Statistical Association*, 10.1080/01621459.2025.2525589.

Mikolov, Tomas (2012) "RNNLM Toolkit," https://www.fit.vut.cz/person/imikolov/public/rnnlm/.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013) "Linguistic Regularities in Continuous Space Word Representations," in Vanderwende, Lucy, Hal Daumé III, and Katrin Kirchhoff eds. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751, Atlanta, Georgia: Association for Computational Linguistics, June, https://aclanthology.org/N13-1090/.

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson (2025) "Sparse Autoencoders for Hypothesis Generation," https://arxiv.org/abs/

2502.04382.

Mullainathan, Sendhil and Ashesh Rambachan (2024) "From Predictive Algorithms to Automatic Generation of Anomalies," Working Paper 32422, National Bureau of Economic Research, 10.3386/w32422.

Nabeshima, Noa (2024) "Matryoshka Sparse Autoencoders," https://www.lesswrong.com/posts/zbebxYCqsryPALh8C/matryoshka-sparse-autoencoders, December, LessWrong blog; published 14 December 2024.

Olshausen, Bruno A. and David J. Field (1996) "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, 381 (6583), 607–609, 10.1038/381607a0.

O'Neill, Charles, Christine Ye, Kartheik Iyer, and John F. Wu (2024) "Disentangling Dense Embeddings with Sparse Autoencoders," https://arxiv.org/abs/2408.00657.

OpenAI, Josh Achiam, Steven Adler et al. (2024) "GPT-4 Technical Report," https://arxiv.org/abs/2303.08774.

Park, Kiho, Yo Joong Choe, and Victor Veitch (2024) "The Linear Representation Hypothesis and the Geometry of Large Language Models," https://arxiv.org/abs/2311.03658.

Patwardhan, Tejal, Rachel Dias, Elizabeth Proehl et al. (2025) "GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks," https://arxiv.org/abs/2510.04374.

Paulo, Gonçalo, Alex Mallen, Caden Juang, and Nora Belrose (2025) "Automatically Interpreting Millions of Features in Large Language Models," https://arxiv.org/abs/2410.13928.

Praveen, G.B., Anita Agrawal, Ponraj Sundaram, and Sanjay Sardesai (2018) "Ischemic stroke lesion segmentation using stacked sparse autoencoder," *Computers in Biology and Medicine*, 99, 38–52, 10.1016/j.compbiomed.2018.05.027.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005) "Teachers, Schools, and Academic Achievement," *Econometrica*, 73 (2), 417–458, http://www.jstor.org/stable/3598793.

Rockoff, Jonah E. (2004) "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94 (2), 247–252, 10.1257/0002828041302244.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger (2011) "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy*, 6 (1), 43–74, 10.1162/edfp_a_00022.

Rose, Evan, Jonathan Schellenberg, and Yotam Shem-Tov (2022) "The Effects of Teacher Quality on Adult Criminal Justice Contact," Working Paper 30274, National Bureau of Economic Research, 10.3386/w30274.

Rothstein, Jesse (2017) "Measuring the Impacts of Teachers: Comment," *American Economic Review*, 107 (6), 1656–1684, 10.1257/aer.20141440.

Sacerdote, Bruce (2011) "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" in *Handbook of the Economics of Education*, 3, Chap. 4, 249–277: Elsevier, JEL code: I2.

Shah, Rajen D. and Richard J. Samworth (2012) "Variable Selection with Error Control: Another Look at Stability Selection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75 (1), 55–80, 10.1111/j.1467-9868.2011.01034.x.

Silverman, B.W. (1998) *Density Estimation for Statistics and Data Analysis*: Routledge, 10.1201/9781315140919.

Slavin, Robert E. (1980) "Cooperative Learning," *Review of Educational Research*, 50 (2), 315–342, 10.3102/00346543050002315.

Staiger, Douglas O. and Jonah E. Rockoff (2010) "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, 24 (3), 97–118, 10.1257/jep.24.3.97.

Stein, Mary Kay, Randi A. Engle, Margaret S. Smith, and Elizabeth K. Hughes (2008) "Orchestrating Productive Mathematical Discussions: Five Practices for Helping Teachers Move Beyond Show and Tell," *Mathematical Thinking and Learning*, 10 (4), 313–340, 10.1080/10986060802229675.

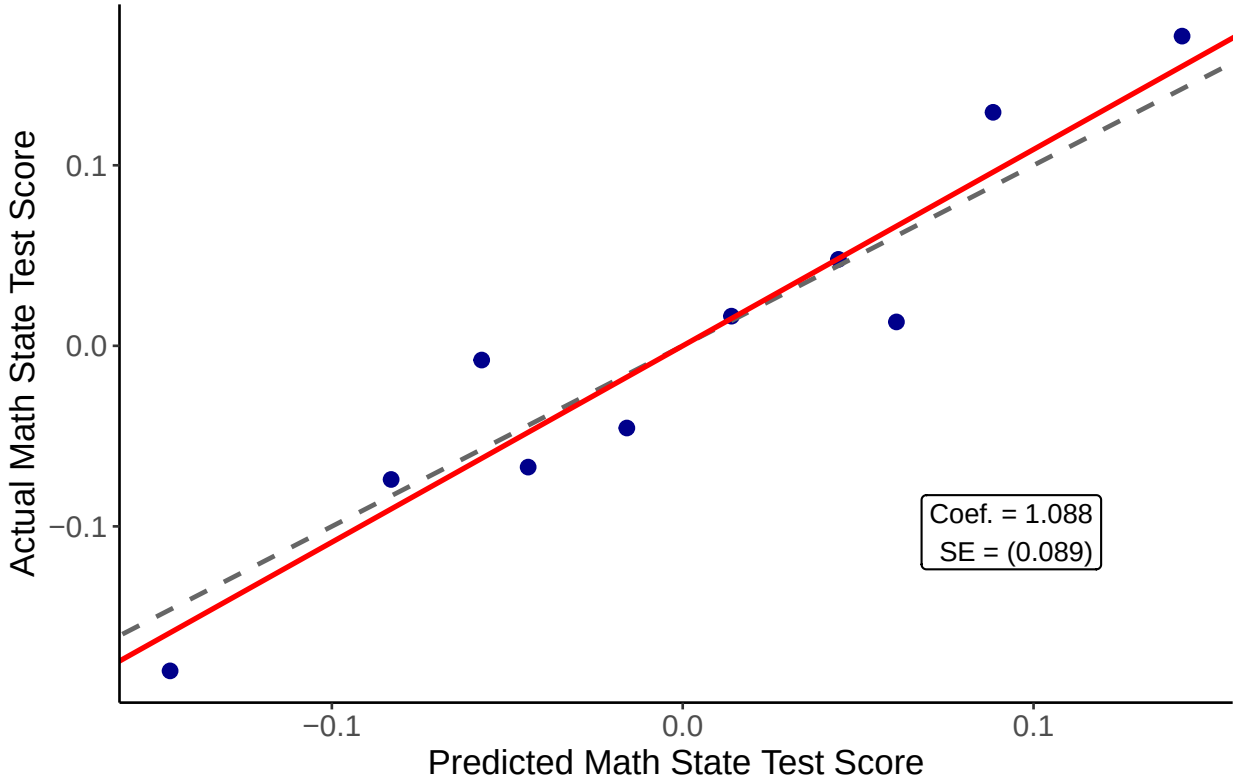
Sun, Wenjun, Siyu Shao, Rui Zhao, Ruqiang Yan, Xingwu Zhang, and Xuefeng Chen (2016) "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, 89, 171–178, 10.1016/j.measurement.2016.04.007.

Templeton, Adly, Tom Conerly, Jonathan Marcus et al. (2024) "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html, May, Transformer Circuits Thread (Anthropic blog); published 21 May 2024.

Zaigrajew, Vladimir, Hubert Baniecki, and Przemyslaw Biecek (2025) "Interpreting CLIP with Hierarchical Sparse Autoencoders," https://arxiv.org/abs/2502.20578.

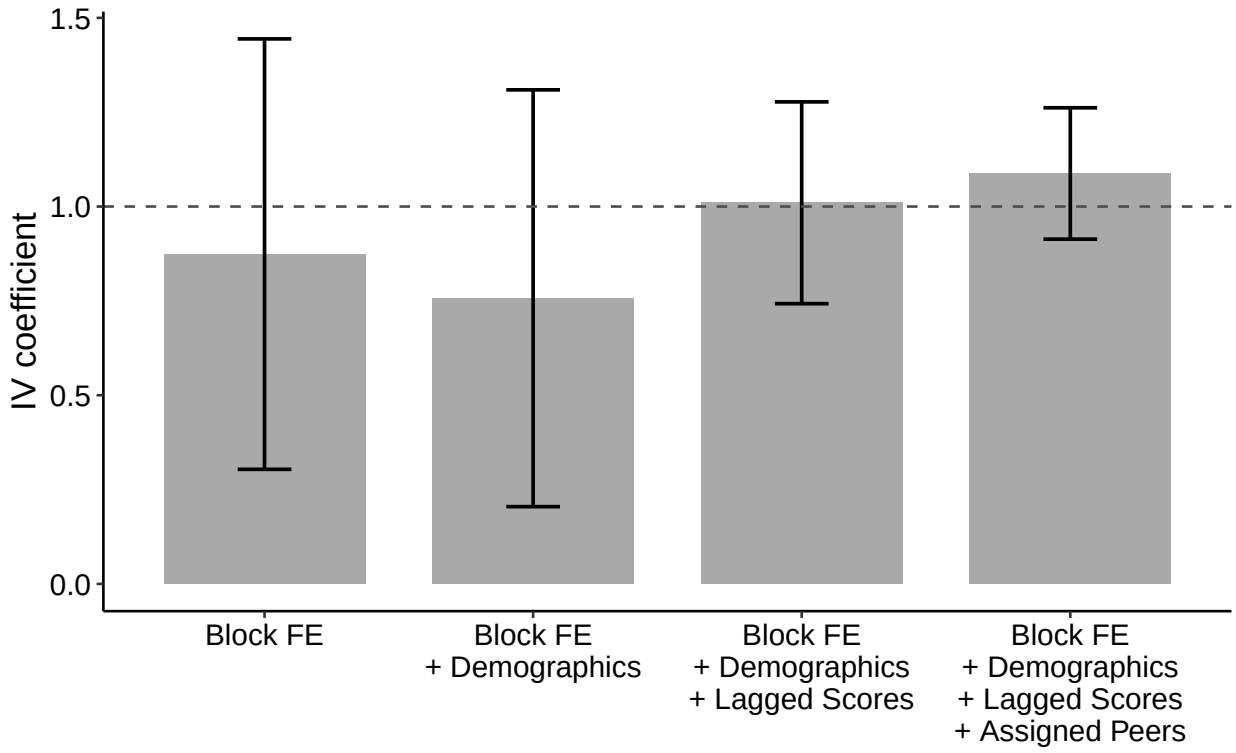
A Additional Empirical Results

Figure A1: Forecast Bias Test: Actual vs. Predicted State Math Test Scores



Note: This figure presents a binned scatterplot of residualized student state math test scores against residualized teacher value-added forecasts formed by taking the inverse-variance-weighted mean of the NCTE-provided shrunken year-specific math value-added estimates from 2009–2012. Both variables are residualized on randomization-block fixed effects, student demographic controls, lagged-test-score controls, and assigned-teacher roster mean classmate characteristics. As described in Section 2.2, value-added forecasts are instrumented by the value-added forecast of the student's randomly assigned teacher. The red line shows the OLS fit on the underlying student-level data, and the dashed line is a 45-degree line. Standard errors are clustered by assigned teacher.

Figure A2: IV Estimates of Value-Added Forecast Calibration Across Specifications



Note: This figure presents IV estimates of the slope coefficient from a regression of student state math test scores on teacher value-added forecasts formed by taking the inverse-variance-weighted mean of the NCTE-provided shrunken year-specific math value-added estimates from 2009–2012. The value-added forecasts are instrumented by the value-added forecast of the student's randomly assigned teacher. All regressions include randomization block fixed effects. Each bar corresponds to a different set of controls: no controls, student demographic variables, student demographic variables and lagged student test scores, and those same controls plus assigned-teacher roster mean classmate characteristics. Error bars represent 95% confidence intervals, with standard errors clustered by assigned teacher. The dashed horizontal line corresponds to a coefficient of one, which is the forecast unbiasedness benchmark.

Table A1: Alternative TVA Prediction Models, Structured Data

	Shrunken TVA			Unshrunken TVA		
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted TVA, Linear Regression	0.091			0.079		
	(0.102)			(0.187)		
Predicted TVA, Ridge		0.402			0.633	
		(0.409)			(0.746)	
Predicted TVA, Gradient Boosted Trees			0.432			1.155
			(0.501)			(0.786)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	261	261	261	261	261	261
R^2	0.003	0.004	0.003	0.001	0.003	0.008
Adjusted R^2	−0.001	0.0001	−0.001	−0.003	−0.0004	0.005

Note: This table reports OLS regressions of teacher value-added on the value-added predictions formed from the structured data using linear regression (Columns 1 and 4), ridge regression (Columns 2 and 5), and gradient boosted trees (Columns 3 and 6). Columns 1 through 3 use shrunken (empirical Bayes) value-added as the dependent variable, while Columns 4 through 6 use unshrunken (fixed effect) value-added. All variables are residualized on school district and grade fixed effects prior to estimation. These regressions are analogous to those reported in Columns 1 and 4 of Table I, except that lasso, rather than the alternative models used here, was used to form the predictions in Table I. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table A2: Alternative TVA Prediction Models, Transcript Embeddings

	Shrunken TVA	Unshrunken TVA
	(1)	(2)
Predicted TVA, Linear Regression	0.159***	0.202**
	(0.061)	(0.102)
Residualized on District & Grade FE	Yes	Yes
Observations	261	261
R^2	0.026	0.015
Adjusted R^2	0.022	0.011

Note: This table reports OLS regressions of teacher value-added on the value-added predictions formed from the transcript embeddings using linear regression. Column 1 uses shrunken (empirical Bayes) value-added as the dependent variable, while Column 2 uses unshrunken (fixed effect) value-added. All variables are residualized on school district and grade fixed effects prior to estimation. These regressions are analogous to those reported in Columns 2 and 5 of Table I, except that ridge regression, rather than linear regression, was used to form the predictions in Table I. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table A3: Alternative TVA Prediction Models, Embedding Algorithms

	Shrunken TVA			Unshrunken TVA		
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted TVA, Gemini Embeddings	1.168***			1.387***		
	(0.218)			(0.406)		
Predicted TVA, Gemma Embeddings		1.345***			1.090*	
		(0.339)			(0.652)	
Predicted TVA, Fine-Tuned Gemma			1.348***			1.298***
			(0.250)			(0.489)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	261	261	261	261	261	261
R^2	0.135	0.059	0.098	0.067	0.014	0.032
Adjusted R^2	0.132	0.055	0.094	0.064	0.010	0.028

Note: This table reports OLS regressions of teacher value-added on the value-added predictions formed from the transcript embeddings using ridge regression with alternative embedding models: Google's Gemini Embedding 2 model, Google's EmbeddingGemma model, and an EmbeddingGemma model fine-tuned as described in Section 3.4. Columns 1 through 3 use shrunken (empirical Bayes) value-added as the dependent variable, while Columns 4 through 6 use unshrunken (fixed effect) value-added. All variables are residualized on school district and grade fixed effects prior to estimation. These regressions are analogous to those reported in Columns 2 and 5 of Table I, except that OpenAI's text-embedding-3-large model was used to form the embeddings in Table I. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table A4: TVA Prediction Using Unshrunken Value-Added Label

	Teacher Value-Added (TVA)			
	(1)	(2)	(3)	(4)
Predicted TVA, Structured Data	0.028			−0.090
	(0.112)			(0.100)
Predicted TVA, Transcript Embeddings (Excluding Year)		0.664*		
		(0.396)		
Predicted TVA, Transcript Embeddings			1.207***	1.252***
			(0.273)	(0.277)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes
Observations	261	185	261	261
R^2	0.0003	0.016	0.089	0.092
Adjusted R^2	−0.004	0.011	0.086	0.085

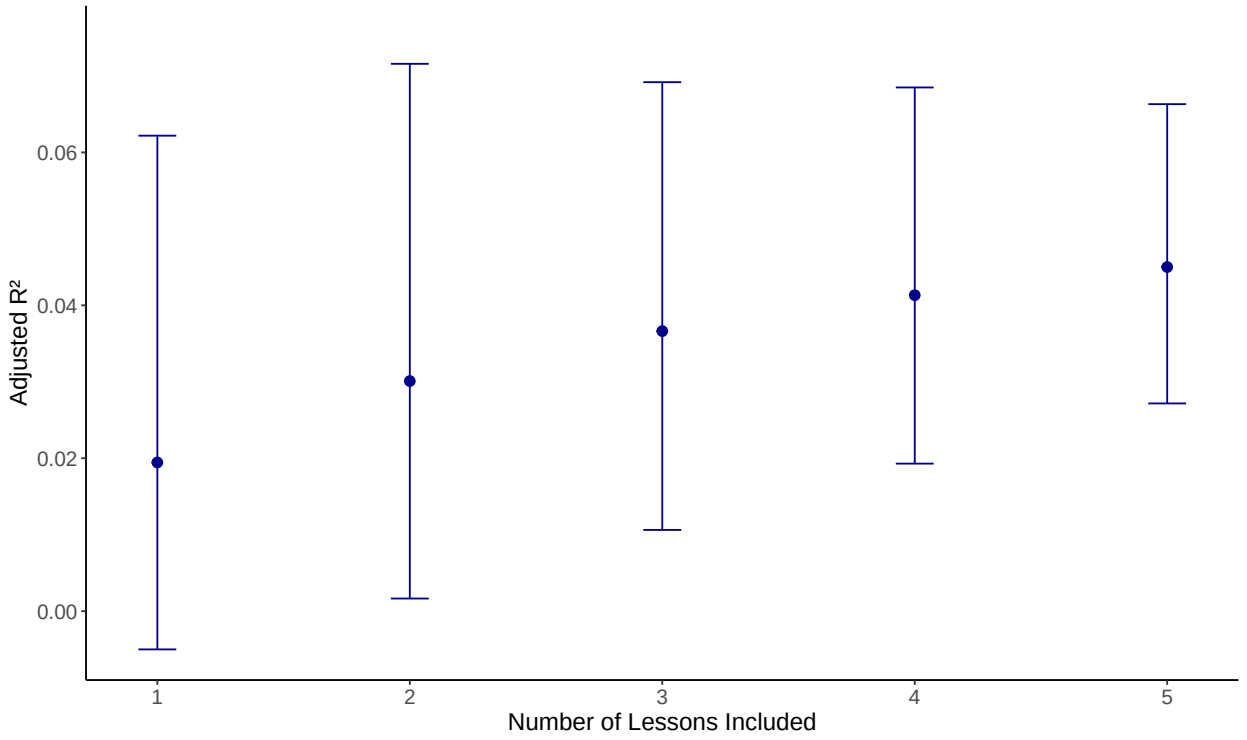
Note: This table reports OLS regressions of teachers' unshrunken (fixed effect) value-added on the predictions formed from the structured data and the transcript embeddings. The predictions in this table are formed using teachers' unshrunken (fixed effect) value-added, residualized on district fixed effects, as the prediction target. Column 1 uses the prediction formed from the structured data using linear regression. Column 2 uses a prediction formed from the transcript embeddings, where the unshrunken value-added estimate being predicted is formed without using student outcome data from the transcript year. Column 3 uses a prediction formed from the transcript embeddings, where the unshrunken value-added estimate being predicted pools data from all available years. Column 4 includes the predictions from Columns 1 and 3 simultaneously. All variables are residualized on school district and grade fixed effects prior to estimation. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table A5: TVA Prediction with Structured Data and Transcripts, Teacher-Year Level

	Shrunken TVA			Unshrunken TVA		
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted TVA, Structured Data	0.024		−0.223	0.019		−0.267
	(0.164)		(0.163)	(0.337)		(0.334)
Predicted TVA, Transcript Embeddings		1.190***	1.277***		1.196***	1.315***
		(0.238)	(0.244)		(0.431)	(0.438)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	895	895	895	496	496	496
R^2	0.0001	0.062	0.066	0.00001	0.019	0.021
Adjusted R^2	−0.001	0.061	0.064	−0.002	0.017	0.017

Note: This table reports OLS regressions of teacher value-added on the value-added predictions formed from the structured data and from the transcript embeddings. The regressions are estimated at the teacher-year level. I consider as possible predictors all of the time-invariant structured information, the teacher's years of experience in that year, and an indicator for whether the teacher is in their first or second year of teaching. I encode teacher experience in this way in light of past research that has found the effect of teacher experience on teacher quality to be very non-linear, with little effect past the first few years (Hanushek 2011). I include all predictors with fewer than 5% missing values across all teacher-year combinations where I observe value-added (the first- or second-year indicator satisfies this condition) and exclude teacher-year observations with missing values for any selected predictor. The labels being predicted are teachers' year-specific shrunken value-added estimates. The transcript embedding predictions are the teacher-level predictions from Table I, repeated across years for each teacher. Columns 1 through 3 use year-specific shrunken (empirical Bayes) value-added as the dependent variable (years 2009–2013), while Columns 4 through 6 use year-specific unshrunken (fixed effect) value-added (years 2011–2013 only). All variables are residualized on school district and grade fixed effects prior to estimation. Standard errors clustered at the teacher level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Figure A3: Adjusted R^2 vs. Number of Transcripts Used for Prediction



Note: The dependent variable is teachers' unshrunken (fixed effect) value-added. This figure visualizes how the adjusted R^2 values of the model specified by Equation (4) (i.e., a regression of teachers' unshrunken value-added on the value-added predictions formed from the transcript embeddings, after residualizing on district and grade fixed effects) change as we vary the number of transcripts used to form the prediction $\hat{y}_k^u$ (i.e., the value-added prediction formed from the transcript embeddings) from one to five. For this exercise, I restrict the sample to all teachers for whom I have at least five transcripts and for whom $\hat{y}_k^s$ (i.e., the value-added prediction formed from the structured data) is non-missing. I randomize the order in which I draw transcripts. When increasing the number of transcripts used to form $\hat{y}_k^u$ from k to $k+1$ (for $k \in \{1, 2, 3, 4\}$), I keep the previous k transcripts and add the additional randomly chosen transcript, rather than randomly reselecting a new set of $k+1$ transcripts. The points correspond to means across 1,000 random transcript orderings. The intervals are 95% confidence intervals formed using these random orderings.

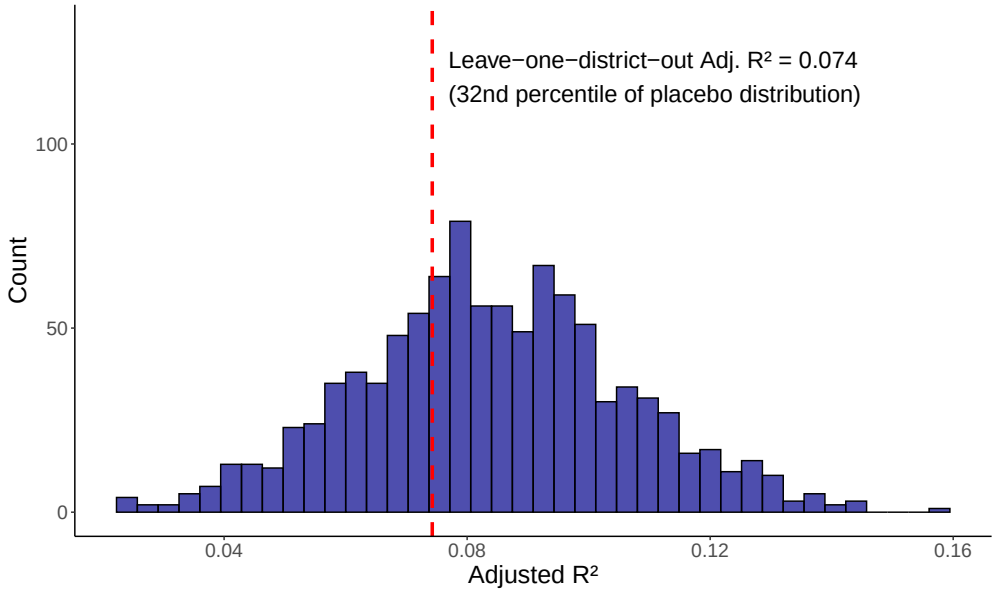
Table A6: Separate TVA Prediction with Student and Teacher Embeddings

	Shrunken TVA		Unshrunken TVA	
	(1)	(2)	(3)	(4)
Predicted TVA from Teacher Embeddings	0.895***		0.978**	
	(0.252)		(0.487)	
Predicted TVA from Student Embeddings		1.235**		1.195
		(0.604)		(1.097)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes
Observations	261	261	261	261
R^2	0.069	0.018	0.029	0.006
Adjusted R^2	0.066	0.014	0.026	0.002

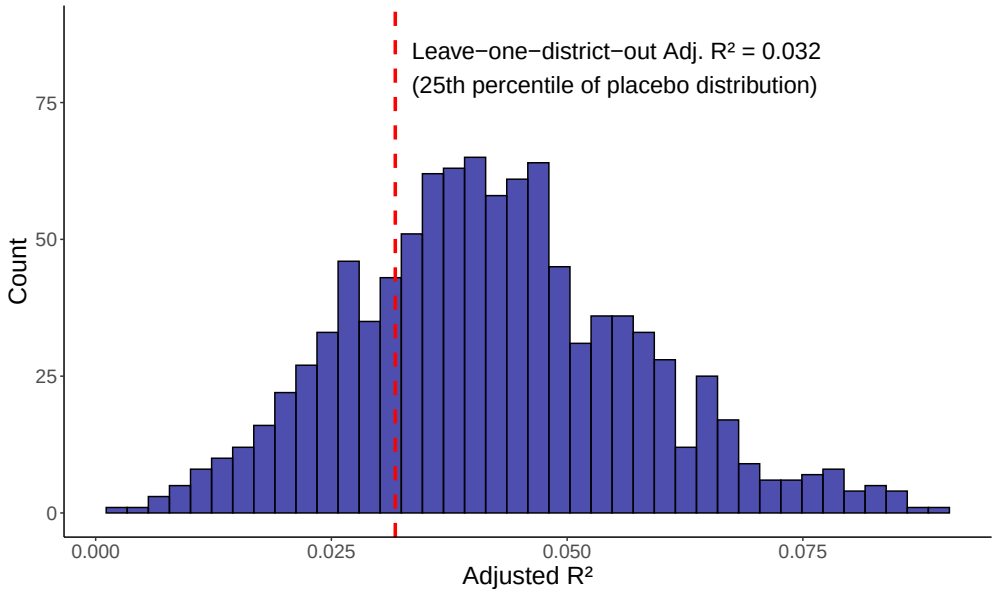
Note: This table reports OLS regressions of teacher value-added on predictions formed using embeddings of only teacher speech (Columns 1 and 3) and only student speech (Columns 2 and 4). Columns 1 and 2 use shrunken (empirical Bayes) value-added as the dependent variable, while Columns 3 and 4 use unshrunken (fixed effect) value-added. All variables are residualized on school district and grade fixed effects prior to estimation. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Figure A4: Leave-One-District-Out Prediction

Panel A: Shrunken Value-Added



Panel B: Unshrunken Value-Added



Note: The dashed red lines indicate the adjusted R^2 values from the model specified by Equation (4) using the leave-one-district-out transcript prediction procedure. The top panel uses teachers' shrunken value-added as the dependent variable, and the bottom panel uses the unshrunken value-added measures. The histograms show the distribution of adjusted R^2 across placebo runs in which teachers are partitioned into random groups matching the size distribution of the actual districts. All variables are residualized on school district and grade fixed effects prior to estimation. The sample is the same as in Table I.

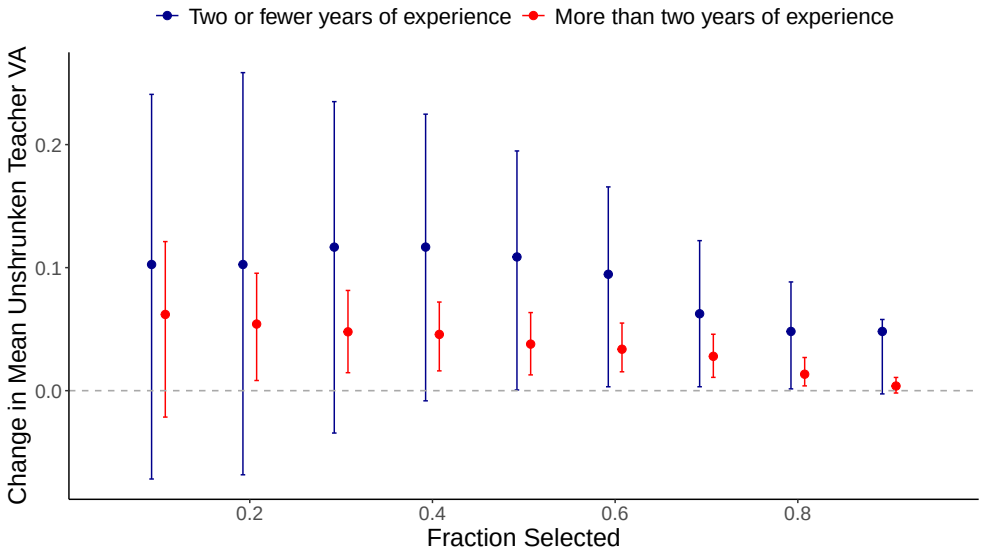
Table A7: TVA Prediction with LLMs

	Shrunken TVA		Unshrunken TVA	
	(1)	(2)	(3)	(4)
Predicted TVA, GPT-5.4	0.084***		0.166***	
	(0.032)		(0.054)	
Predicted TVA, Gemma 4		0.062*		0.139***
		(0.032)		(0.051)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes
Observations	261	261	261	261
R^2	0.027	0.017	0.037	0.031
Adjusted R^2	0.023	0.014	0.034	0.027

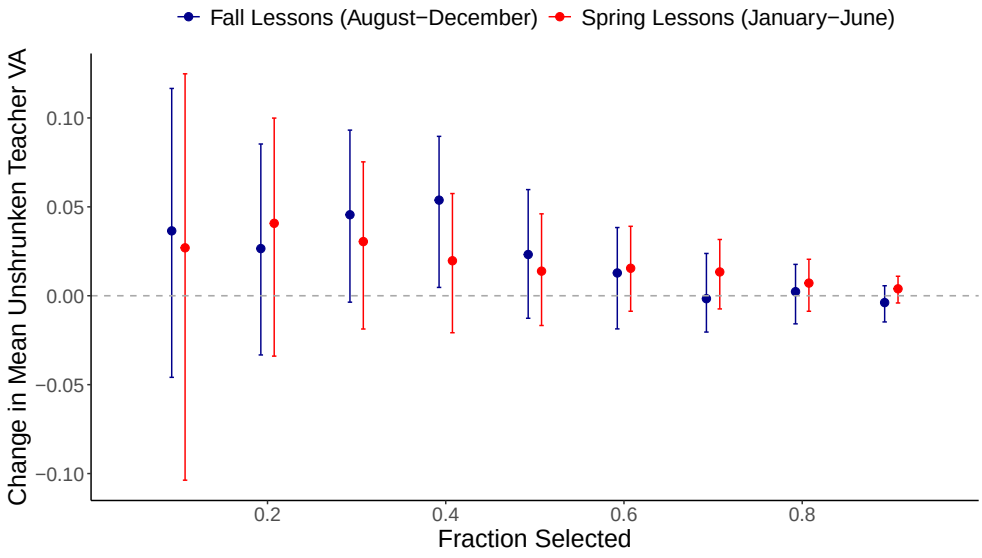
Note: This table reports the results of OLS regressions of teachers' value-added on the value-added predictions made by GPT-5.4 (Columns 1 and 3) and Gemma 4 (Columns 2 and 4) given teachers' classroom transcripts. Columns 1 and 2 use shrunken (empirical Bayes) value-added as the dependent variable, while Columns 3 and 4 use unshrunken (fixed effect) value-added. All variables are residualized on school district and grade fixed effects prior to estimation. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Figure A5: Teacher Hiring Gains by Teacher Experience and Timing of Lesson

Panel A: By Teacher Experience



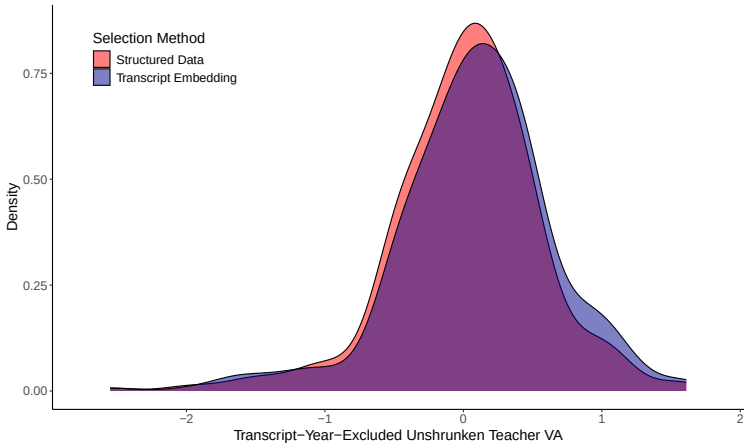
Panel B: By Lesson Month



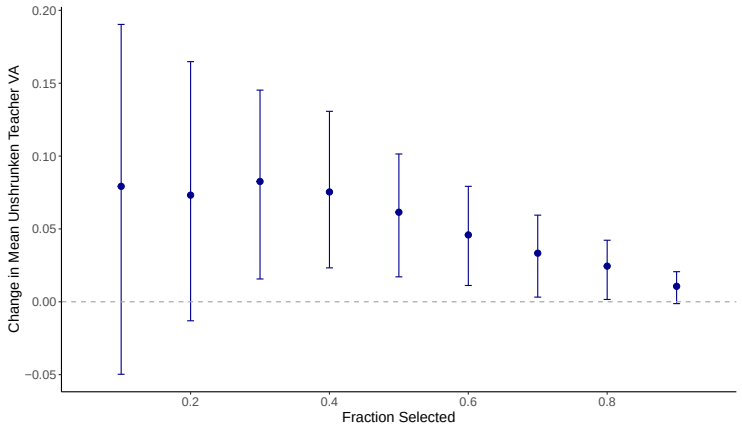
Note: This figure presents estimates of the increase in mean unshrunken value-added resulting from using the transcript predictions, rather than the structured data predictions, to choose teachers within each district, for different fractions of teachers hired. Panel A presents these estimates separately by teacher experience. Panel B presents these estimates separately by whether the randomly chosen lesson occurred in the fall or spring, restricting the sample to teachers with lessons in both parts of the school year. The 95% conditional bootstrap confidence intervals are based on 1,000 within-district bootstrap replicates and 200 lesson draws per replicate, holding the fitted prediction models fixed.

Figure A6: Teacher Hiring, Excluding the Transcript Year from the Outcome

Panel A: Value-Added Distributions Under Two Teacher Selection Policies



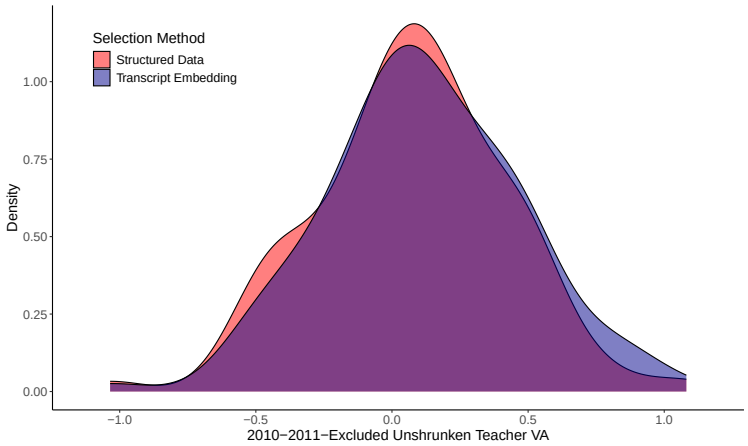
Panel B: Increase in Mean Value-Added Relative to Selection Using Structured Data



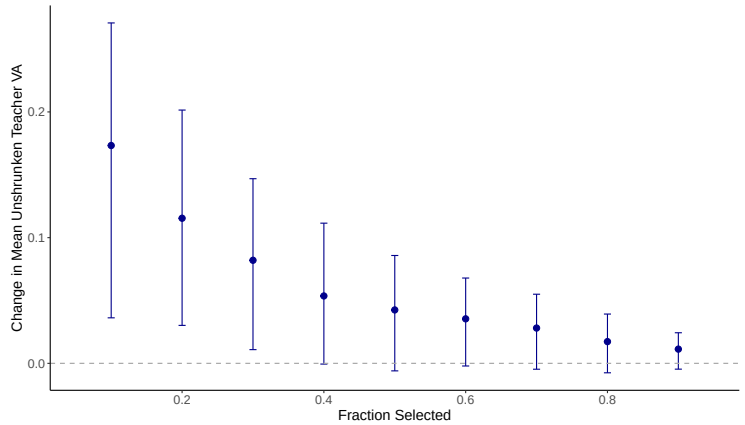
Note: Panel A displays kernel density estimates for two value-added distributions: one for teachers in the top half of predicted value-added within each district based on the structured data (red distribution) and another for teachers in the top half within each district based on transcript data (blue distribution). The density estimates are generated using a Gaussian kernel with bandwidth selection following the method provided by Silverman (1998). Panel B presents estimates of the increase in mean unshrunken value-added resulting from using the transcript predictions, rather than the structured data predictions, to choose teachers within each district, for different fractions of teachers hired (as described in Section 4.1). In both panels, the value-added outcome is unshrunken value-added estimated using years other than the year of the randomly chosen lesson. The 95% conditional bootstrap confidence intervals are based on 1,000 within-district bootstrap replicates and 200 lesson draws per replicate, holding the fitted prediction models fixed.

Figure A7: Teacher Hiring, First-Year Lessons and Excluding the First Year from the Outcome

Panel A: Value-Added Distributions Under Two Teacher Selection Policies

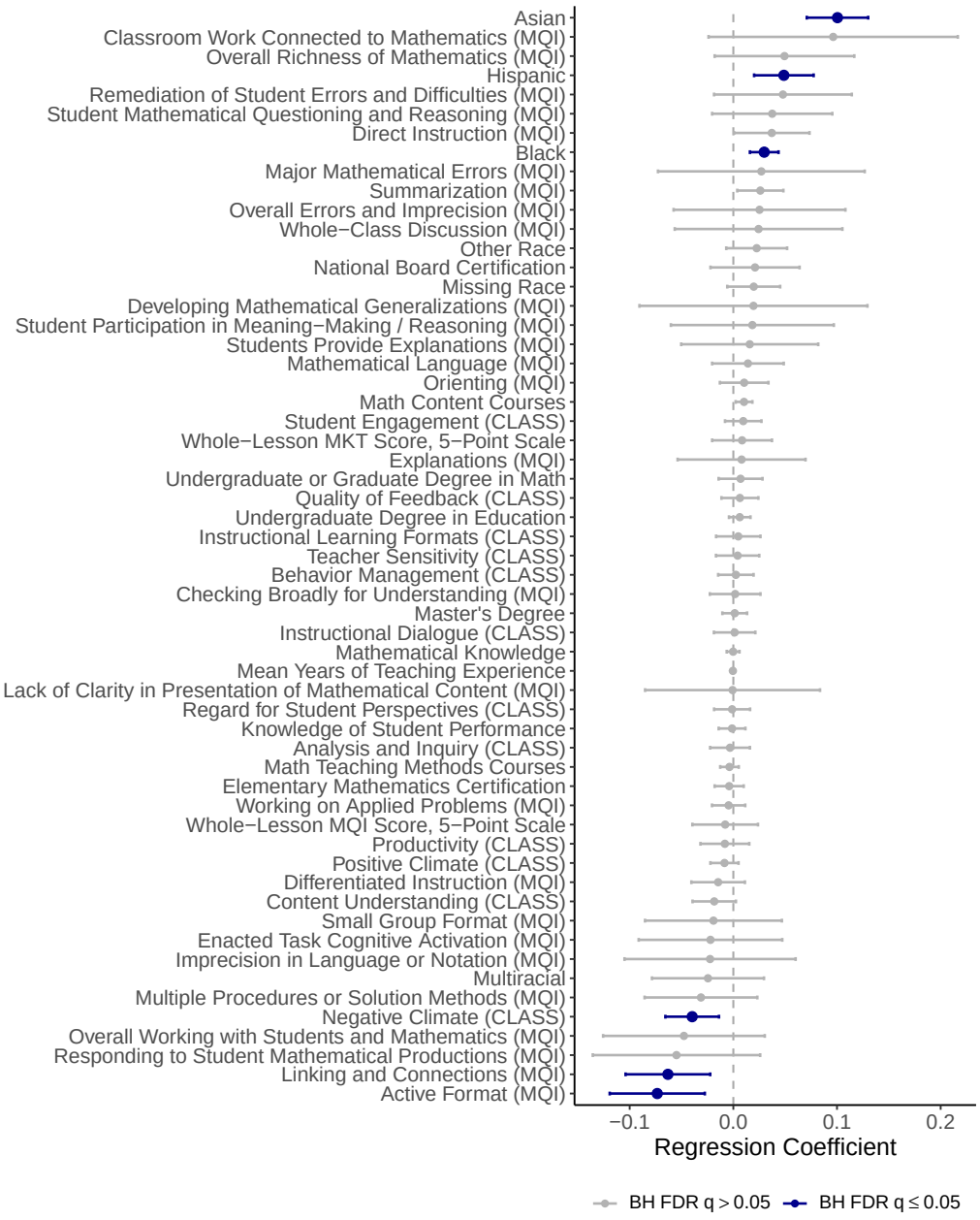


Panel B: Increase in Mean Value-Added Relative to Selection Using Structured Data



Note: Panel A displays kernel density estimates for two value-added distributions: one for teachers in the top half of predicted value-added within each district based on the structured data (red distribution) and another for teachers in the top half within each district based on transcript data (blue distribution). The density estimates are generated using a Gaussian kernel with bandwidth selection following the method provided by Silverman (1998). Panel B presents estimates of the increase in mean unshrunken value-added resulting from using transcript predictions formed from first-year lessons, rather than structured data predictions, to choose teachers within each district, for different fractions of teachers hired (as described in Section 4.1). In both panels, the value-added outcome is unshrunken value-added estimated using only the last two NCTE Main Study years, excluding the first year used to form the transcript predictions. The 95% conditional bootstrap confidence intervals are based on 1,000 within-district bootstrap replicates and 200 lesson draws per replicate, holding the fitted prediction models fixed.

Figure A8: Relationship Between Predicted TVA from Transcripts and Teacher Characteristics



Note: This figure presents coefficient estimates from a regression of predicted value-added from transcripts on the set of "structured" predictors, estimated jointly in a single regression. Coefficients that remain statistically significant after a Benjamini–Hochberg false discovery rate correction at $q \leq 0.05$ are shown in dark blue; the remaining coefficients are shown in gray. Error bars represent 95% confidence intervals. The White race dummy and the fraction of time spent in a mix of active and small group instruction are omitted to avoid perfect collinearity.

Table A8: TVA Prediction with Transcript Embeddings, Race and Gender Controls

	Shrunken TVA	Unshrunken TVA
	(1)	(2)
Predicted TVA, Transcript Embeddings	1.102***	1.369***
	(0.239)	(0.436)
Residualized on District & Grade FE	Yes	Yes
Residualized on Race & Gender	Yes	Yes
Observations	257	257
R^2	0.081	0.044
Adjusted R^2	0.078	0.040

Note: This table reports OLS regressions of teacher value-added on the value-added predictions formed from the transcript embeddings, residualizing all variables on teacher race and gender indicators in addition to school district and grade fixed effects prior to estimation. Column 1 uses shrunken (empirical Bayes) value-added as the dependent variable, while Column 2 uses unshrunken (fixed effect) value-added. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table A9: TVA Prediction with Leave-One-Teacher-Out IPSS-Selected Sparse Autoencoder Features

	Shrunken TVA		Unshrunken TVA	
	Main Sample	Full Sample	Main Sample	Full Sample
	(1)	(2)	(3)	(4)
Predicted value-added	0.628***	0.644***	0.855**	0.893***
	(0.177)	(0.171)	(0.342)	(0.332)
Residualized on District & Grade FE	Yes	Yes	Yes	Yes
Observations	261	294	261	294
R^2	0.043	0.043	0.028	0.028
Adjusted R^2	0.039	0.040	0.025	0.025

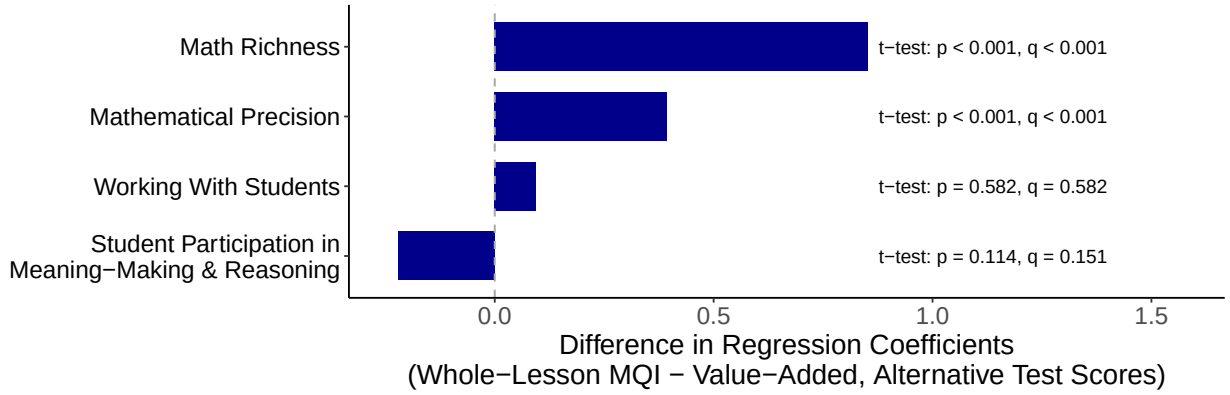
Note: This table reports OLS regressions of teacher value-added on out-of-sample predictions formed from IPSS-selected teacher-level SAE features, as described in Section 5.3. Predictions are formed by leaving out one teacher at a time, running IPSS on the remaining teachers using teacher-level SAE activations as features and shrunken teacher value-added as the outcome, fitting a post-selection linear model on the selected features, and predicting value-added for the held-out teacher. For each leave-one-teacher-out split, IPSS uses lasso as the base selector with 250 penalty values, 500 complementary half-sample draws, and the cubic IPSS transformation. Columns 1 and 2 use shrunken value-added as the dependent variable, while Columns 3 and 4 use unshrunken value-added as the dependent variable. Within each outcome group, the first column restricts to the main prediction sample used in Table I, while the second column uses all teachers for whom both transcripts and value-added data are available. All variables are residualized on school district and grade fixed effects prior to estimation. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table A10: Relationship Between Transcript Predictions and IPSS-Selected Features

	Predicted TVA					
	(1)	(2)	(3)	(4)	(5)	(6)
Feature 73: mentions sitting or chairs	−0.485***					−0.482***
	(0.020)					(0.020)
Feature 850: mentions the organization of students into specific teams or groups		0.283***				0.291***
		(0.040)				(0.041)
Feature 854: mentions the instruction to "switch"			0.195***			0.196***
			(0.034)			(0.034)
Feature 1016: mentions a mnemonic involving McDonald's and cheeseburgers				−0.013		−0.019
				(0.052)		(0.052)
Feature 680: teacher explicitly invites a student to come up to the front of the classroom or to the board					−0.215***	−0.200***
					(0.021)	(0.021)
Observations	74,516	74,516	74,516	74,516	74,516	74,516
R^2	0.012	0.002	0.001	0.00000	0.002	0.017
Adjusted R^2	0.012	0.002	0.001	−0.00001	0.002	0.017

Note: This table reports OLS regressions of value-added predictions formed from the small (approximately 200 token) transcript segments used in Section 5 on each IPSS-selected SAE cluster activation feature. The predictions are formed by applying the model fit on larger transcript embeddings in Section 3 to transcript segments from the held-out teacher. These regressions are estimated at the transcript segment level. Standard errors, reported in parentheses, are clustered at the teacher level.

Figure A9: Differences Between Whole-Lesson MQI and Alternative Value-Added Coefficients



Note: This figure plots the difference between the coefficient from a regression of teachers' whole-lesson MQI scores on the instructional feature and the coefficient from a regression of teachers' unshrunken alternative value-added scores on the same instructional feature. Positive values indicate that the instructional feature is more strongly associated with whole-lesson MQI scores than with value-added on the end-of-year mathematics tests designed and administered by the NCTE Main Study team. The instructional features are indices constructed using principal component analysis on the MQI rubric items, as described in Section 2.3. The alternative value-added scores are estimates of value-added on scores from an assessment designed by the NCTE Main Study researchers to measure higher-level mathematics skills that may be underrepresented on state exams. The two dependent variables and the instructional features are standardized to have mean zero and standard deviation one. All instructional features are included in the regressions together with district fixed effects, as described in Section 6. The p-values come from regressions of standardized whole-lesson MQI minus standardized alternative value-added on all four MQI instructional features and district fixed effects. The Benjamini-Hochberg q-values adjust these p-values for multiple testing across the four instructional features.

B Additional Prediction Materials

B.1 Zero-Shot LLM Prediction Prompt

The following prompt is used to elicit predictions of teacher effectiveness from GPT 5.4 and Gemma 4, as described in Section 3.4.

```
Below is a transcript from inside an elementary school math classroom.
   I would like you to generate a numerical rating from 1 to 5
   representing your assessment of how effective the teacher will be at
    raising students' state math test scores, where 1 is extremely
   ineffective, 3 is average, and 5 is extremely effective. Please
   adhere to this rating system and avoid inflating or deflating your
   rating. Please only output a number (either an integer or, if needed
   , a float) between 1 and 5, without any additional text. When
   forming your judgment, please consider everything you know about
   what makes an effective teacher and what you think a classroom with
   effective teaching looks like (including student behavior, if
   applicable). It is very important that you complete this task to the
    best of your ability, so please ensure that your assessment
   reflects your best judgment.
```

C Feature Generation and Selection

C.1 Sparse Autoencoder (SAE) Architecture and Training Procedure

C.1.1 Notation and Problem Setup

Let $X \in \mathbb{R}^D$ denote a single input embedding of dimension $D = 3072$. For a given SAE seed, let $W_e \in \mathbb{R}^{M \times D}$ denote encoder weights, $W_d \in \mathbb{R}^{D \times M}$ denote decoder weights, $b_x \in \mathbb{R}^D$ denote the input bias, and $b_h \in \mathbb{R}^M$ denote the hidden bias.

In my implementation, the Matryoshka dictionary sizes are given by

$$\mathcal{M} = \{16,\ 32,\ 64,\ 128,\ 256,\ 512\}.$$

At most $K = 16$ hidden units can be active on every forward pass. Moreover, during training, the data are processed in mini-batches of size $B = 1024$. Additionally, let $X^{(j)}$ denote the j-th sample in the current mini-batch.

The activation function is the rectified linear unit (ReLU) $\sigma(z) = \max\{0, z\}$. Also define

$$\mathrm{TopK}(u, K) = \{\, i : u_i \text{ is among the } K \text{ largest components of } u \,\}.$$

C.1.2 Encoder, Sparsification, and Decoder

The input is first centered and mapped to pre-activations $u \in \mathbb{R}^M$:

$$\widetilde{X} = X - b_x$$

$$u = W_e \widetilde{X} + b_h.$$

The activation, $h \in \mathbb{R}^M$ is given by

$$h_i = \begin{cases} \sigma(u_i) & \text{if } i \in \mathrm{TopK}(u, K), \\ 0 & \text{otherwise.} \end{cases}$$

For every $m \in \mathcal{M} = \{16, 32, 64, 128, 256, 512\}$, define

$$h_i^{(m)} = \begin{cases} h_i & \text{if } i \le m, \\ 0 & \text{if } i > m, \end{cases}$$

and

$$\widehat{X}^{(m)} = W_d\, h^{(m)} + b_x.$$

When the full dictionary ($m = 512$) is used, I abbreviate $\widehat{X} \equiv \widehat{X}^{(512)}$.

C.1.3 Loss Function

Define the batch-normalized mean-squared error

$$\ell_{\text{norm}}(\widehat{X}, X) = \frac{\frac{1}{B}\sum_{j=1}^{B} \|\widehat{X}^{(j)} - X^{(j)}\|_2^2}{\frac{1}{B}\sum_{j=1}^{B} \|X^{(j)} - \bar{X}\|_2^2 + \varepsilon},$$

where B is the mini-batch size, $\bar{X} = \frac{1}{B}\sum_{j=1}^{B} X^{(j)}$, and $\varepsilon = 10^{-9}$ prevents division by zero.

In addition to the Matryoshka reconstruction objective, I use an auxiliary loss intended to prevent features from becoming permanently inactive during training. For each mini-batch sample j, let $\mathcal{D}$ denote the set of "dead" neurons, defined as neurons with activation $h_i = 0$ for more than 256 consecutive mini-batches. Among these dead neurons, define $A^{(j)}$ as the set of the $K_{\text{aux}} = 32$ largest pre-activations for sample j, and construct

$$h^{(j)}_{\text{aux},i} = \begin{cases} \sigma(u_i^{(j)}) & \text{if } i \in A^{(j)}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $R^{(j)} = X^{(j)} - \widehat{X}^{(j)}$ denote the main reconstruction residual and let

$$\widehat{R}^{(j)}_{\text{aux}} = W_d h^{(j)}_{\text{aux}}$$

denote the reconstruction of this residual using only the selected dead neurons. I define the auxiliary loss as

$$\ell_{\text{aux}} = \ell_{\text{norm}}(\widehat{R}_{\text{aux}}, R),$$

with gradients from this term flowing through W_d and the auxiliary activations but not through the main residual R.

Let α denote the auxiliary-loss weight, which I set to $\frac{1}{32}$ in my implementation. The mini-batch loss is

$$\mathcal{L} = \frac{1}{|\mathcal{M}|}\sum_{m \in \mathcal{M}} \ell_{\text{norm}}\big(\widehat{X}^{(m)}, X\big) + \alpha\, \ell_{\text{aux}}.$$

C.1.4 Training Procedure

I implement the SAE training procedure using the HypotheSAEs package (Movva et al. 2025). I train 10 SAEs with random seeds 1001 through 1010. Each training run uses all available transcript-segment embeddings, runs for 50 epochs, and uses a learning rate η of 5×10^{-4}. The input bias b_x is initialized to the component-wise median of the embeddings, the columns of W_d are drawn from a Xavier-uniform distribution and then ℓ_2-normalized, W_e is initialized to $W_d^\top$, and b_h is set to zero.

For each mini-batch, I compute the mini-batch loss $\mathcal{L}$ as defined above and backpropagate to obtain the gradient $\nabla_\theta \mathcal{L}$. I then remove the component of the decoder gradient parallel to each decoder column, clip the global gradient norm to at most 1.0, and perform a single update with step size η using the Adam optimizer (Kingma and Ba 2017). Finally, to enforce the decoder's unit-norm constraint, I renormalize each column of the weight matrix W_d to have an ℓ_2-norm of 1.

C.2 Clustering Features Across SAE Seeds

Let $s \in \{1001, \ldots, 1010\}$ index SAE training seeds, and let $h_{is}(X)$ denote the activation of feature i from seed s on embedding X. I cluster the 10×512 learned features using agglomerative clustering.

For two learned features a and b, I define the distance

$$d(a, b) = \frac{1}{2}\left(1 - \cos(w_a, w_b)\right) + \frac{1}{2}\left(1 - \max\{\rho(h_a, h_b), 0\}\right),$$

where w_a and w_b are the normalized decoder vectors and $\rho(h_a, h_b)$ is the correlation between the two features' activations on a random sample of transcript segments. Thus, two features are treated as similar when both their learned directions and their patterns of activation across transcript segments are similar. I use average-linkage agglomerative clustering and cut the dendrogram at a distance threshold of 0.35.

For a cluster c, let $\mathcal{G}_c$ denote the set of seed-feature pairs assigned to that cluster. I define the cluster activation for embedding X as

$$A_c(X) = \frac{1}{|\mathcal{G}_c|} \sum_{(s,i) \in \mathcal{G}_c} h_{is}(X),$$

and define the cluster presence indicator as

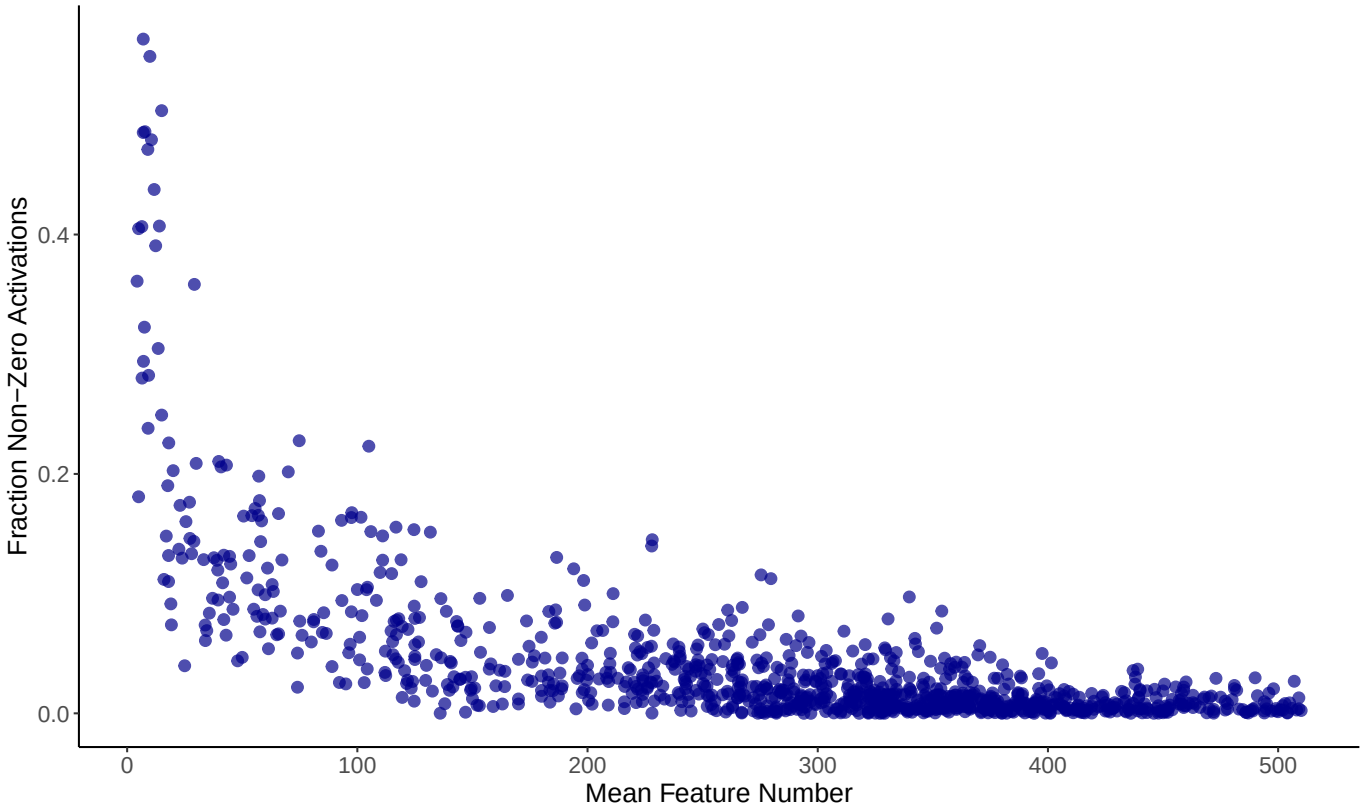
$$P_c(X) = \mathbf{1}\left\{ \max_{(s,i)\in\mathcal{G}_c} h_{is}(X) > 0 \right\}.$$

The teacher-level SAE feature used in downstream analyses is the mean of $A_c(X)$ across that teacher's transcript segments. I refer to each resulting cluster by the cluster feature number assigned by the clustering algorithm.

C.3 Feature Prevalence

Appendix Figure C1 reports, for each cluster feature, the share of transcript segments where at least one member SAE feature is active. Features are indexed by the average position of their member features in the underlying SAE dictionaries. Activation rates decrease sharply as the average feature number increases. This is a consequence of the Matryoshka objective: low-average-index features must be able to by themselves form relatively accurate reconstructions and therefore tend to be broadly useful across segments, whereas later features specialize in narrower phenomena that are used only when the larger dictionary is available.

Figure C1: Prevalence of Cluster Features



Note: This figure displays the prevalence of the cluster features constructed from the multi-seed Sparse Autoencoder procedure described in Appendix C. It plots, for each cluster feature, the share of transcript segments where at least one member SAE feature has a non-zero activation. Formally, this is the mean of the cluster presence indicator $P_c(X)$ defined above.

C.4 Integrated Path Stability Selection

Integrated path stability selection (IPSS; Melikechi and Miller 2025) builds on stability selection by applying a base variable-selection algorithm repeatedly to complementary subsamples and tracking how frequently each feature is selected. Rather than evaluating selection frequencies at a single penalty value, IPSS integrates information over the regularization path.

IPSS controls the number of expected false positives (EFP) under the assumptions that the observations are i.i.d. and that features unrelated to value-added are not systematically selected across repeated random splits of the data beyond what would be implied by the average sparsity of the base selector, when averaged over the regularization path. Under these assumptions, IPSS provides an EFP score for each feature, defined as the smallest bound on the expected number of false positives under which that feature would be selected. IPSS then converts these scores into q-values that approximate the false discovery rate by comparing the EFP bound with the number of selected features.

For the primary feature-selection exercise, I use lasso as the base selector with 500 penalty values, draw 2,500 complementary half-samples of teachers, and use the cubic IPSS transformation. I use shrunken value-added as the outcome variable and retain features with q-values below 0.20.

For the leave-one-teacher-out prediction exercise reported in Appendix Table A9, I rerun IPSS separately within each training sample. To reduce the computational burden of this nested procedure, I use 250 penalty values and 500 complementary half-samples in each split, while otherwise retaining the same specification.

C.5 LLM Feature Interpretation and Validation

For each IPSS-selected cluster, I select 10 transcript segments with the highest cluster activations and randomly choose 10 transcript segments where no member feature in the cluster is active. I provide these positive- and non-activating samples to the open-source LLM Gemma 4 and ask it to generate a natural-language description by identifying the most specific attribute present in the highly activating samples but absent from the non-activating samples.

I evaluate each generated description on a separate scoring sample. For each selected cluster, I draw 100 new positive-activating segments and 100 non-activating segments and ask the LLM whether the feature described in the interpretation is present in each segment. I then compare the resulting annotations with the SAE cluster presence indicator using the F1 score.

Let i index the 200 scoring samples, let P denote the indices of the positive-activating samples, let N denote the indices of the non-activating samples, and let a_i denote the binary LLM annotation for sample i. Define

$$S_P = \sum_{i \in P} a_i \qquad \text{and} \qquad S_N = \sum_{i \in N} a_i.$$

Recall is

$$\frac{S_P}{100},$$

and precision is

$$\frac{S_P}{S_P + S_N}.$$

The F1 score is

$$\frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}.$$

C.5.1 Feature Interpretation

This prompt is based on the HypotheSAEs prompt, with task-specific instructions added.

```
You are a machine learning researcher who has trained a neural network
   on a text dataset. You are trying to understand what text features
   cause a specific neuron in the neural network to fire.

You are given two sets of SAMPLES: POSITIVE SAMPLES that strongly
   activate the neuron, and NEGATIVE SAMPLES from the same distribution
    that do not activate the neuron.
```

```
Your goal is to identify a feature that is present in the positive
   samples but absent in the negative samples.
The texts are elementary and middle-school mathematics classroom
   transcript chunks.
Look for concrete instructional, conversational, or classroom-
   management patterns that distinguish the positive samples from the
   negative samples.

POSITIVE SAMPLES:
----------------
{positive_texts}
----------------

NEGATIVE SAMPLES:
----------------
{negative_texts}
----------------

Rules about the feature you identify:
- The feature should be objective, focusing on concrete attributes
   rather than abstract concepts.
- The feature should be present in the positive samples and absent in
   the negative samples. Do not output a generic feature which also
   appears in negative samples.
- The feature should be as specific as possible, while still applying
   to all of the positive samples. For example, if all of the positive
   samples mention Golden or Labrador retrievers, then the feature
   should be "mentions retriever dogs", not "mentions dogs" or "
   mentions Golden retrievers".

Do not output anything else. Your response should be formatted exactly
   as shown in the examples above.
Please suggest exactly one feature, starting with "-" and surrounded by
    quotes "". Your response is:
- "
```

C.5.2 Interpretation Scoring

This prompt comes from the HypotheSAEs package.

```
Check whether the TEXT satisfies a PROPERTY. Respond with "Yes" or "No
   ". When uncertain, output "No". Do not include any other text in
   your response.

PROPERTY: "{hypothesis}"
TEXT: "{text}"
Output ("Yes" or "No" only):
```

C.6 Examples of IPSS-Selected Features Predictive of Value-Added

Table C1: Examples of IPSS-Selected SAE Features

Feature Number	Top Activating Examples
73	1. teacher: So right now you guys need to sit up. Everybody, sit up including you. Push in your chairs. Push in your chair. Fix your chair. I know it's hot. I know you just came from lunch but it's not an excuse for you to space out. Sit properly, feet under your desk, Student S, same with you. Same with you. Sit properly, your back against your chair. All right and before we get into our big activity, or lesson I should say for the day because it's not really an activity—so yesterday we did questions that had one scale. 2. teacher: Front row, right there. student: [Inaudible] the content. teacher: Exactly, exactly. Those are, like, prime seats right there. Yeah, right in the corner. teacher: Okay. teacher: Okay, okay. All set? All right, so let's do a little arranging. Yeah, so you know what? Bring a chair over. You want to sit over here in the chair? student: [Inaudible]. teacher: Okay. Student A, are you all set? student: [Inaudible]. 3. teacher: Stand up. Take that chair and sit right there. Great job. How did you decide that it was there? student: 'Cause it's five-tenths. teacher: It's five-tenths. But how did you decide as a group that that's – this particular spot right there is where it should go? student: Because we started, um, where [inaudible] and did 10, 20, 30. teacher: So you just counted where one-tenth was and you went one-tenth, two-tenths, three-tenths, four-tenths, five-tenths? student: Fifty percent. teacher: One and twenty-five hundredths. You need to have a seat. student: Can I do one of these? 4. student: He said I cheated because – teacher: No, that's easy. That's good. teacher: Go sit down. teacher: All right, 3, 2, everyone to your seat. Do not come to me. Go to your seat. I'm coming to you. Go to your seat. 5. teacher: You can [inaudible]. Have a seat. student: [Inaudible]. teacher: Yes, you can join. [Dinging].
850	1. teacher: Team 2 is getting y'all. 64. 6 to the fourth power, 6 to the fourth power. 6 to the fourth power. I don't see where you've calculated the answer. You're guessing. Did you talk to anybody? student: [Inaudible] teacher: Team 2 got it again. student: [Inaudible]

Continued on next page

Table C1 – *Continued*

Feature Number	Top Activating Examples
	2. teacher: Team three. Anything left? Just tell me pass if you don't have anything left. It's okay. student: Pass. teacher: Okay. One minute. Team four? student: Exponent. teacher: Exponent. Team one. student: Multiply. teacher: Okay, we have multiply. Subtraction? student: We have that. teacher: We have subtract? Okay. It's got to be something we don't have a form of. student: Expression. teacher: Expression. Okay, team two. student: 90 degrees. teacher: 90 degrees. Team three. student: Pass. teacher: Team four. Equation? [Laughter] 3. student: Contiguous. teacher: Contiguous. Okay. If you used that, mark it off. Team two. student: Multiply. teacher: Multiply. Team three. student: Corner. teacher: Corner. We'll just see how many words we come up with. Team four. student: Parenthesis. teacher: Parenthesis. Team one, do you have another word? student: Order of operations. teacher: Very good, order of operations. It's okay if you run out. Just tell me you pass, okay? Team two? student: Linking. teacher: Linking. Team three. student: Subtract. teacher: Subtract. Team four. student: Pattern. teacher: Pattern. Excellent. Team one. student: Addition. teacher: Addition. Team two. student: Division. 4. teacher: Okay. Decide where you want your tile. The first team's going to have to decide what color they are. multiple students: [Multiple comments]. teacher: Okay guys? Can you wait for a moment before you start another one? Listen. It doesn't make a lot of sense for the two people on the opposing team to just be sitting there. Okay? One of you does need to be watching the timer. The other one could be writing out expressions, using those same numbers to check the opposing team to see if you come up with any of the same values, right?

Continued on next page

Table C1 – *Continued*

Feature Number	Top Activating Examples
	5. teacher: So, right now, just before we get started, I need you all to get in two teams. I'm gonna tell you what to do. I'm gonna have one team here, and one team here. Let me count how many students are here today. 2, 4, 6, 8, 10, 12, 14, 15. Ooh, so it's a little off. We have one extra person. So, what we'll do is, one team will have seven, one team will have eight. Okay? So, the team it could work it either way, because the team that has eight could have an advantage because they'll have an extra person that could help them out. The team that has seven may be at an advantage because they'll be without a person, so in case someone else is not too sure – so, it could work out either way. Student T? It could work out either way. Okay? So, here are your groups. All right, first person ____. Can you grab a chair, grab your chair and come up here please?
854	1. teacher: Just make your guess. Switch! Switch! Switch. student: What's so funny? teacher: I'm just looking at discussion, switch. Trying to get my answer, Student J? Switch. student: [Inaudible]. teacher: I like that. I'm good at that. student: Nobody's [Inaudible]. teacher: Favorite president, switch. student: [Inaudible]. teacher: Good. Don't tell me what to pick. Please tell me one thing about President Garfield. student: He was the fattest president. teacher: Good. Yeah, he was like the cat. student: [Inaudible]. teacher: Do you really know that? student: Yeah. teacher: Really? student: I watched both movies of [Inaudible]. So do I. He knows [Inaudible]. teacher: And you don't? student: No. teacher: Really? 2. teacher: Oh, R & B, yeah. Well, I said between five and four and six, I said 'cause like. Like if you didn't – for the fifth grade teacher, that'd be a good question but you can't make up other teacher. Switch. student: What's a BRB? [Inaudible]. teacher: It doesn't matter, Student L, just answer it. Just answer it. student: All these just pick on something. Or I can guess something. Guys [Inaudible]. teacher: Switch. Just, just answer the question. Quit telling every, switch. All right, switch. Now, this is your last one. student: [Inaudible] put Wayne for [Inaudible]. I put number one. Mr. W? teacher: Switch. student: Mr. W, just let you know I put five or I put four.

Continued on next page

Table C1 – *Continued*

Feature Number	Top Activating Examples
	3. teacher: Student E, where you going? student: I'll switch with [inaudible]. teacher: When I ask you to get up and switch, that's when I want you to switch. student: Okay. teacher: Are we ready? multiple students: Yes. teacher: Student E, Student E hasn't had anybody to switch with. student: I'll switch with you. teacher: You don't have anybody to switch with? student: No, I do. I do. 4. teacher: Yes. We cannot give it to our friends and we don't know the answer, so let's first solve it in our heads and then we're gonna pass it on, 'cause we need to know if they're right or wrong, right? All right, get busy. Should take you maybe a minute and you switch. [Long pause] teacher: You did it? Oh, you switched already? student: No. teacher: No, okay. student: Are we supposed to do one? 5. teacher: Mm-hmm. Student J will be the writer and Student S will be the writer. It don't work that way this time. So when I say switch, whoever is writing needs to pass that pen to the next person. And when I say switch again, whoever is writing needs to pass it on to the next person. But everybody else needs to be doing some work while that person is writing. Correct? multiple students: Yes. teacher: We got it? multiple students: Yes. teacher: One more time. Student J, what are we looking for? student: Givens. teacher: What are the givens? student: That five children want to get three slices.
1016	1. teacher: And how Mc Donald's wanted to what – what would you want it to do? student: You want it to sell more burgers. teacher: Remember the Mc Donald's case where you were the CEO and you decided you wanted to build a Mc Donald's in China? This is the second part to our problem. The new Mc Donald's case. You ready? multiple students: Yes. 2. teacher: There's no remainder, so McDonald's does not sell cheeseburgers raw today. McDonald's does not sell cheeseburgers raw today. Excuse me, Teacher D, I'm taping – Teacher D, excuse me, I'm taping right now. Thank you. That's okay, things happen. It's the end of the year. We forget. Where was I? We just realized McDonald's doesn't sell cheeseburgers raw. Why does McDonald's not sell cheeseburgers raw today? teacher: There isn't a remainder. That could be a good thing because McDonald's don't want to get anyone sick. All right, so another way we can do this. Now, we can go back. We can check, right? We have an answer right here. Who remembers what we call the answer when we're doing division? What is his name? Student H?

Continued on next page

Table C1 – *Continued*

Feature Number	Top Activating Examples
	3. teacher: But if it's something that we can remember, that's engaging to you little people, hey, I don't mind trying it. So, divide, multiply, subtract, compare, sometimes a remainder. I always take each letter from this, and I come up with something different. And I usually say, sometimes, "Does McDonald's sell cheeseburgers?" That's a question. Does McDonald's sell cheeseburgers? And I could always ask, "What about do they sell them raw?" Because sometimes we need that remainder. So, that's a silly little saying so that you can remember the steps, D, M, S, C, sometimes R. Does McDonald's sell cheeseburgers raw? We have to divide, multiply, subtract, compare, and sometimes a remainder. Yes? student: My [inaudible] teacher, she does just "Does McDonald's sell cheeseburgers." 4. teacher: I'm glad you guys are all – I love your enthusiasm. We just have to make sure we're working within the rules. So after we multiply this, you're going to subtract it just like you always do. 39 minus 36 is 3, and what are we going to do with the next digit here, 8? We've come up with our little helper to make sure we understand all the skills here in long division, and that is Mc Donald's serves burgers. We've multiplied and got a 36. That's the M. We subtracted 36 from 39. That's the S, and then we brought down the 8. That's the B, bring down. So we're going to start the process all over again. You have a question? student: [Inaudible]. 5. teacher: So we could still do that with long division, right? Does 4 go into 1? Remember, does Mc Donald's serve cheeseburgers? Duh, then we check it. Does 4 go into 1? multiple students: No. teacher: So what is it? multiple students: 0. teacher: Then I multiply. What's 4 times 0? multiple students: 0. teacher: What property is that? multiple students: 0. teacher: 4 minus 0, Student C. student: 4 minus 0? 1. teacher: This is the part where everyone's getting stuck. Do what? multiple students: [Inaudible]. teacher: Or burgers. Whatever you like best. So now I have a 1 and I bring down the 0. Does 4 go into 10? multiple students: Yes. multiple students: No.

Continued on next page

Table C1 – *Continued*

Feature Number	Top Activating Examples
680	1. student: I multiplied the – teacher: Bring down. Bring down. Fabulous, Student T. Okay, let's look at the front. Turn your papers over. Is there anybody that would like to – anybody that would like to come up? Oh, I have a few volunteers. Okay. teacher: Student L, would you like to come up? You are. No? student: Pass. teacher: Okay, Student J, would you like to come up? student: No. student: I didn't do anything on the front. teacher: You didn't do anything on the front, on this side? student: No, I didn't get it. teacher: I didn't see your hand up. Student J, would you like to come up? student: Uh. teacher: Student J? Student L? student: Can I just go up. teacher: Student C? Would you like to come up? 2. teacher: What we're gonna do is I have a sheet. You're gonna use the information to help collect some data about classmates alright. Um, I actually need a volunteer. Um, Student T come on up. you're actually gonna be working in pairs when doing this. can you come stand right over here for me? 3. teacher: Okay. If you can get this drawn out, I would like you to come up and show it. Perfect. Would you be willing to come up and show that? What do you put there? Okay. Would anyone like to come up and show what they came up with for an answer, and I can move my stuff out of the way, and move this, and turn it back. Feeling brave? Student M, come on up. Remember to speak up a little bit so we can hear you. You can sit in the chair if you want. What did you do up here, Mr. M? student: I drew a square and cut it into thirds. I shaded in two of them. Then I drew another square and subtracted it. teacher: Okay. SO you drew another square and subtracted 1? Got it. 4. teacher: I was going across. All right, let's look at the next one. We're gonna do a couple more. up here and do the next one for me. First I'm going to write the problem down, and then I need someone to come up and show me. 5 and 2-ninths plus 3 and 1-ninth. Someone come up here. Student J. Talk to us. student: [Inaudible].

Continued on next page

Table C1 – *Continued*

Feature Number	Top Activating Examples
	5. student: I came up with the 4th floor. teacher: I'm going to ask Student D to come up and draw what she did, because I think this will give you a very great visual. Okay. Come on up and show me how you solved your problem. What I want you to do as an audience to be respectful, please put the caps on the dry erase markers, place the dry erase markers in front of you. Have body control. Your eyes are on the presenter, because if you were up there presenting, you would want respect from your classmates. Student D, do you mind talking while you're doing it? What are you doing? student: [Inaudible].

Note: This table reports the top activating transcript-segment examples used to interpret each IPSS-selected SAE feature.

C.7 Stability-Selected Features Predictive of Whole-Lesson MQI Scores

Table C2: IPSS-Selected SAE Features for Whole-Lesson MQI

Feature Number	Correlation with Whole-Lesson MQI	Correlation with Residualized Unshrunken TVA	IPSS q-value	EFP Score	Description	F1 Score
271	0.289	0.111	0.050	0.078	the teacher explicitly uses the word "prove" or "proof" to instruct students to justify their mathematical answers	0.942
1007	-0.314	-0.081	0.050	0.102	contains teacher singing or singing nursery rhymes	N/A
687	0.145	-0.021	0.050	0.151	mentions the word "net" or "nets" in the context of 3D shapes being flattened	0.880
614	0.186	0.066	0.053	0.254	mentions the substitution of specific letters (like F or S) for numerical values in equations or expressions	0.131

Continued on next page

Table C2 – *Continued*

Feature Number	Correlation with Whole-Lesson MQI	Correlation with Residualized Unshrunken TVA	IPSS q-value	EFP Score	Description	F1 Score
1066	0.238	-0.021	0.053	0.265	teacher explains the conceptual reasoning or the "why" behind a mathematical process or rule	0.383
339	0.346	0.138	0.112	0.974	discussions or questions regarding the definition or naming of a "strategy"	0.450
894	0.314	0.046	0.112	1.023	teacher discusses alternative or more efficient mathematical methods or procedures	0.554
85	0.341	0.119	0.112	1.061	mentions the specific phrases "what do you notice" or "what do you observe"	0.630
801	0.239	0.019	0.112	1.094	mentions the word "pieces"	0.970
993	-0.322	0.011	0.112	1.118	mentions specific geographic locations, celebrities, or famous landmarks/attractions	0.434
102	-0.067	0.064	0.134	1.475	mentions the act of choosing or making a choice	0.985

Note: This table presents the SAE features selected using the IPSS procedure described in Section 5 to predict teachers' whole-lesson MQI scores. The first column reports the cluster feature number. The second column reports the Pearson correlations between teacher-level mean activations and teachers' whole-lesson MQI scores. The third column reports the correlation between teacher-level mean activations and unshrunken teacher value-added after residualizing value-added on school district and grade fixed effects. The fourth and fifth columns report the IPSS q-value and expected false-positive score used for feature selection. The sixth column reports the natural-language description of the feature generated by having the open-source LLM Gemma 4 compare transcript segments where the feature is present to those where the feature is not present. Finally, the seventh column reports the F1 score measuring how well the natural-language description distinguishes activating from non-activating transcript segments, as described in Section 5.3. The score ranges from 0 to 1, with higher values indicating better agreement between the description-based LLM ratings and whether the SAE feature is active in the segment. For Feature 1007, the F1 score is reported as N/A because the LLM did not classify any scored segments as positive, making precision and the F1 score undefined.

D Additional Details on Data Construction

D.1 Teacher Value-Added

D.1.1 Value-Added Model Components

In the model

$$a_{i,j,k,t} = A_{i,t-1}^{\top}\alpha + S_{i,t}^{\top}\beta + P_{j,k,t}^{\top}\delta + E_{i,t}^{\top}\rho + \mu_k + \theta_{k,t} + \epsilon_{i,j,k,t} \tag{1}$$

$a_{i,j,k,t}$ represents the rank-based mathematics state test score for student i in class j taught by teacher k during school year t, standardized at the district, grade, and subject level. $A_{i,t-1}$ captures prior student performance. It includes a cubic polynomial of student i's mathematics state test score $a_{i,t-1}$ in the previous school year, an interaction term between the prior score and the grade level of the current test, and student i's previous-year score $a'_{i,t-1}$ on the state English exam. $S_{i,t}$ includes the following student characteristics: gender, race and ethnicity, eligibility for free or reduced-price lunch, English Language Learner status, and special education designation. $P_{j,k,t}$ represents peer influences and includes the means of the characteristics in $S_{i,t}$ among students in class j, the means of $a_{i,t-1}$ and $a'_{i,t-1}$ among students in class j, the fraction of students in class j missing scores for $a_{i,t-1}$ and $a'_{i,t-1}$, and the number of students in class j. $P_{j,k,t}$ also includes the same variables measured at the school-grade level, where the number of students in class j is replaced by the number of students in the grade. $E_{i,t}$ contains fixed effects for student i's district and for the grade-year cell in which student i takes the test in year t. μ_k and $\theta_{k,t}$ capture teacher and teacher-year random effects, respectively, and $\epsilon_{i,j,k,t}$ is an idiosyncratic error term.

D.1.2 Comparison of Value-Added Estimates Across Samples

This section compares the shrunken value-added estimates provided by Kane et al. (2015) with those obtained by estimating Equation (1) using the data available to me, which cover only the 2010–2011, 2011–2012, and 2012–2013 school years and only include students whose teachers participated in the study.

We observe that the two sets of estimates are highly correlated, with a Pearson correlation coefficient of 0.861. Moreover, Appendix Table D1 shows that a regression of the NCTE-provided estimates on the ones I compute yields a coefficient statistically indistinguishable from 1 (the point estimate is 0.976 with a standard error of 0.035).

Table D1: NCTE-Provided vs. Restricted-Sample Shrunken Value-Added

	NCTE-Provided shrunken value-added
Restricted-sample shrunken estimate	0.976***
	(0.035)
Constant	0.006
	(0.004)
Observations	309
R^2	0.741
Adjusted R^2	0.740

Note: This table reports the estimates from an OLS regression of the NCTE-provided shrunken value-added estimates on shrunken estimates computed by estimating Equation (1) on the restricted sample available to me. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

D.1.3 Forecast Bias Test

Let $a_{i,j,k,2013}$ denote student i's standardized state math test score in the random-assignment year, and let $\widehat{\mu}_k^{-2013}$ denote the inverse-variance-weighted mean of the NCTE-provided shrunken year-specific math value-added estimates from 2009–2012.

If students perfectly complied with their teacher assignments, one could test for forecast unbiasedness by testing whether the slope τ in the regression

$$a_{i,j,k,2013} = \tau \widehat{\mu}_{k(i)}^{-2013} + C_i^{\top}\gamma + \lambda_{b(i)} + \varepsilon_i \tag{6}$$

is equal to one, where $k(i)$ is student i's actual teacher, $\lambda_{b(i)}$ are randomization-block fixed effects, and C_i is an optional vector of controls. To account for noncompliance, I estimate τ using two-stage least squares (2SLS), instrumenting the effectiveness of a student's actual teacher with the effectiveness of their randomly assigned teacher. The first stage regression is

$$\widehat{\mu}_{k(i)}^{-2013} = \pi \widehat{\mu}_{k^A(i)}^{-2013} + C_i^{\top}\kappa + \lambda_{b(i)} + \nu_i, \tag{7}$$

and the second-stage regression is (6), where $\widehat{\mu}_{k(i)}^{-2013}$ is replaced by its fitted value from (7). The estimation sample consists of 61 assigned teachers and 827 students. The first-stage regression has an F-statistic of 4,827.3, indicating that the instrument is sufficiently strong.

In Appendix Figure A1, I visualize the second-stage regression with a binned scatterplot, using prior student performance $A_{i,t-1}$, student demographics $S_{i,t}$, and the mean

characteristics of the students assigned to the same teacher as the control variables C_i. The estimated $\hat{\tau}$ is 1.09 with a standard error of 0.09, meaning that we cannot reject the null of forecast unbiasedness. Appendix Figure A2 shows the estimated $\hat{\tau}$ and corresponding 95% confidence interval obtained across specifications with no controls, student demographics, student demographics and prior student performance, and those same controls plus assigned-teacher roster mean classmate characteristics. The confidence intervals include one across all specifications.

D.2 Selected Structured Predictors

The structured prediction model uses the following 59 teacher characteristics. Except for mean years of teaching experience, each has fewer than 5% missing observations among teachers with value-added scores: Math Content Courses, Math Teaching Methods Courses, Undergraduate or Graduate Degree in Math, Undergraduate Degree in Education, Elementary Mathematics Certification, Master's Degree, National Board Certification, Hispanic, Black, White, Asian, Multiracial, Other Race, Missing Race, Positive Climate (CLASS), Negative Climate (CLASS), Teacher Sensitivity (CLASS), Regard for Student Perspectives (CLASS), Behavior Management (CLASS), Productivity (CLASS), Instructional Learning Formats (CLASS), Content Understanding (CLASS), Analysis and Inquiry (CLASS), Quality of Feedback (CLASS), Instructional Dialogue (CLASS), Student Engagement (CLASS), Active Format (MQI), Both Active and Small Group Format (MQI), Small Group Format (MQI), Direct Instruction (MQI), Working on Applied Problems (MQI), Classroom Work Connected to Mathematics (MQI), Whole-Class Discussion (MQI), Linking and Connections (MQI), Explanations (MQI), Multiple Procedures or Solution Methods (MQI), Developing Mathematical Generalizations (MQI), Mathematical Language (MQI), Overall Richness of Mathematics (MQI), Remediation of Student Errors and Difficulties (MQI), Responding to Student Mathematical Productions (MQI), Overall Working with Students and Mathematics (MQI), Major Mathematical Errors (MQI), Imprecision in Language or Notation (MQI), Lack of Clarity in Presentation of Mathematical Content (MQI), Overall Errors and Imprecision (MQI), Students Provide Explanations (MQI), Student Mathematical Questioning and Reasoning (MQI), Enacted Task Cognitive Activation (MQI), Student Participation in Meaning-Making / Reasoning (MQI), Orienting (MQI), Summarization (MQI), Checking Broadly for Understanding (MQI), Differentiated Instruction (MQI), Whole-Lesson MQI Score, 5-Point Scale, Whole-Lesson MKT Score, 5-Point Scale, Mathematical Knowledge, Knowledge of Student Performance, Mean Years of Teaching Experience.

D.3 Summary Statistics

Table D2: Summary Statistics of Selected Teacher Characteristics

Statistic	Mean	St. Dev.	Min	Max
Math Content Courses	1.478	0.831	0	3
Math Teaching Methods Courses	1.338	0.775	0	3
Undergraduate or Graduate Degree in Math	0.058	0.233	0	1
Undergraduate Degree in Education	0.532	0.500	0	1
Elementary Mathematics Certification	0.155	0.362	0	1
Master's Degree	0.766	0.424	0	1
National Board Certification	0.014	0.119	0	1
Mean Years of Teaching Experience	10.388	7.193	0.000	36.500
Hispanic	0.029	0.167	0	1
Black	0.212	0.410	0	1
White	0.651	0.477	0	1
Asian	0.029	0.167	0	1
Multiracial	0.011	0.104	0	1
Other Race	0.025	0.157	0	1
Missing Race	0.043	0.204	0	1
Male	0.168	0.374	0	1
Whole-Lesson MQI Score, 5-Point Scale	2.943	0.471	1.667	4.500
Shrunken Value-Added on State Math Test Scores	0.007	0.148	−0.475	0.446
Unshrunken Value-Added on State Math Test Scores	−0.025	0.269	−0.946	0.683

Note: This table presents summary statistics for a selected set of teacher characteristics for the 278 teachers for whom all of these variables are non-missing. I include teacher characteristics with fewer than 5% missing values among the set of teachers with value-added scores, and I additionally include teachers' mean years of experience. I do not report summary statistics for each classroom observation category but instead report summary statistics for the whole-lesson MQI score, which reflects raters' assessment of the overall quality of the lesson. Math Content Courses and Math Teaching Methods Courses are coded such that a value of zero indicates that the teacher took no such courses, a value of one indicates that the teacher took one or two such courses, a value of two indicates that the teacher took three to five such courses, and a value of three indicates that the teacher took at least six such courses.

D.4 Indices of Instructional Practices

Table D3: Classroom Observation Variables Used in Index Construction

Index Name	**Instrument**	**Component Variables**
Emotional Support	CLASS	Positive Climate, Teacher Sensitivity, Regard for Student Perspectives
Instructional Support	CLASS	Instructional Learning Formats, Content Understanding, Analysis and Inquiry, Quality of Feedback, Instructional Dialogue
Classroom Organization	CLASS	Negative Climate, Behavior Management, Productivity
Student Engagement	CLASS	Student Engagement
Math Richness	MQI	Linking and Connections, Explanations, Multiple Procedures or Solution Methods, Developing Mathematical Generalizations, Mathematical Language
Working with Students	MQI	Remediation of Student Errors and Difficulties, Responding to Student Mathematical Productions in Instruction
Errors and Imprecision	MQI	Major Mathematical Errors, Imprecision in Language or Notation (Mathematical Symbols), Lack of Clarity in Presentation of Mathematical Content
Student Participation in Meaning-Making and Reasoning	MQI	Students Provide Explanations, Student Mathematical Questioning and Reasoning, Enacted Task Cognitive Activation

Note: This table presents the classroom observation variables used to construct the indices used in Section 6.

D.5 Example Classroom Transcript

Figure D1: Example Transcript Excerpt

speaker	text
teacher	Friends, yesterday we started off by working on some word problems together, yes?
student	Yes.
teacher	And yesterday towards the end of the period, you got a word problem given to you that you had to draw a picture to show it, and you had to do the math problem to show it, yes?
multiple students	Yes.
teacher	Some of you might be done. If you are finished, then you can just look over it to make sure that you are happy with the way that it is. And if you are not done, now is your time to get it done. I'm gonna set about ten minutes on my timer. And if we need less time than that, then we will take less time. But I'm gonna give you at least ten minutes to finish it up. And then we're gonna start something new. So my timer is set for ten minutes. You get ten minutes to finish it up.
student	Can we find new ways to check it or [inaudible]?
teacher	Well, the best way to check it is going to be to draw the picture and do the math part of it. So let's look at what you have. A lawn mower holds three fourths a gallon of gas. So I'd be drawing that picture. I could be drawing a whole gallon cut into fourths with three of it filled.
student	Can you [find] the gallon?
student	I don't know how to.
teacher	I guess draw it – basically remember yesterday when I showed you the cups, like a cup of cheese. You're just drawing a container and cut into fourths because you need to have three-fourths of it filled. So that's the whole amount.
student	I can make it like that. That's the way it fills up the gas.
teacher	That works. So now show three fourths of that.
student	I'm pretty [inaudible].
teacher	You might wanna draw it a little bit bigger. Like I wanna show this. Like yesterday when we were talking about the cheese I showed you a picture. And I said that she needed three fourths. So that's one fourth, two fourths, three fourths, four fourths would be the whole thing. And I said that she had to have this much, right?
student	Mm-hmm.
teacher	Do the same thing with what you're doing there but with a gallon of gas. I'm talking to you just in case. Draw a picture of how much she would have in all with the new color. Basically a picture of your end result. Let me see what you got going on here.
student	He's using my pencil.
teacher	That's what you have going on? Because Student H was using your pencil?
student	Yeah.
teacher	So four, five.
student	At first [inaudible].
teacher	I'm writing it upside down. Hold on. Let's turn it around [inaudible]. There we go.
student	[Inaudible]. And then she erased it.
teacher	So basically Student J trying to get to the last house, right?
student	Mm-hmm.
teacher	And how far is it, the whole distance?
student	Four fifths a mile.
teacher	Four fifths a mile. And she's already walked half. And you're trying to find how much further she has to go. So you say that four fifths is your total and that she's already done half, right? So you jot down your LCM is ten. So you know it is tenths. So your answer is three tenths of a mile. So let me see your picture. So you need to label your picture. So you're still working on it.
student	Student J here. Student I here. And then four fifths in the middle and put a line here.
student	No, we put right here or right in the middle.

Note: This figure presents the first 30 rows of the NCTE Main Study utterance-level transcript data as prepared by Demszky and Hill (2023).

D.6 Human Value-Added Predictions

In this section, I explain how I construct the dataset of human value-added predictions used in Section 6.

I observe teacher identification numbers for the 27 teachers included in the value-added prediction exercise described in Section 6, which allows me to link these teachers with their characteristics and value-added scores.[47] In addition, I observe the mapping between estimated value-added ranks and predicted value-added ranks within each district via Figure 2 of Hill et al. (2013), which is reproduced here as Appendix Figure D2.[48] In Hill et al. (2013), school districts are identified with letters (B, G, and R); I map these onto the district numbers observed in the NCTE Main Study data based on the data provided in Table 1 of Hill et al. (2013) on student demographics disaggregated by school district. District B corresponds to District 11, District G corresponds to District 14, and District R corresponds to District 12.

Given that I observe teachers' value-added scores and the mapping from estimated value-added rankings to predicted value-added rankings within each district, I should in principle be able to reconstruct the within-district value-added rankings used by Hill et al. (2013) and use Appendix Figure D2 to identify the predicted value-added rank for each teacher. However, the analysis conducted in Hill et al. (2013) used value-added scores constructed with data only through the 2011–2012 school year, whereas the value-added estimates I observe make use of an additional year of data.

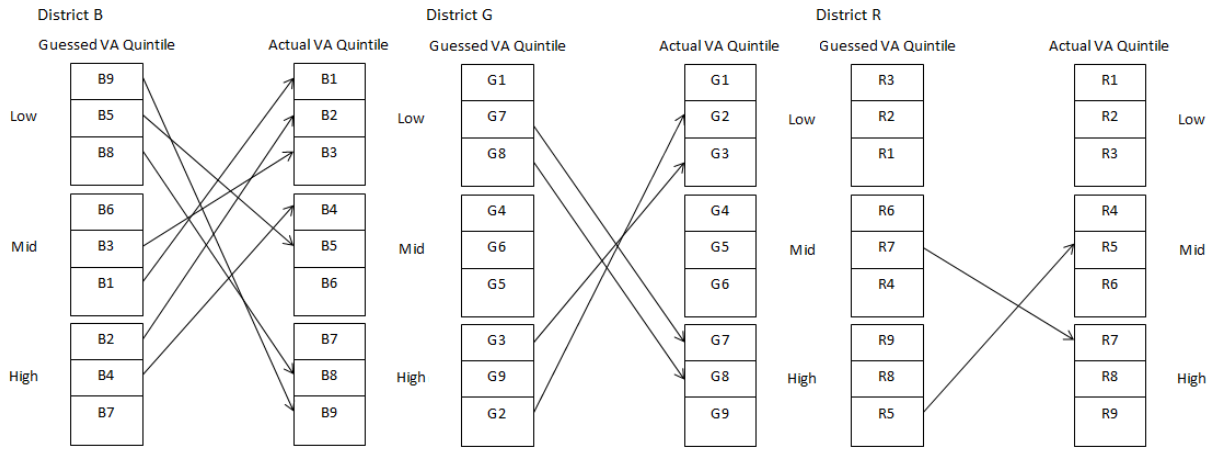
I can, however, closely approximate the value-added scores used in Hill et al. (2013) by computing the simple mean of the NCTE-provided year-specific shrunken value-added estimates through the 2011–2012 school year. These means need not line up exactly with the shrunken teacher value-added estimates one would obtain by estimating Equation (1) using data through the 2011–2012 school year, but they are close approximations. For example, if one regresses the NCTE-provided all-years shrunken value-added estimate on the simple mean of the year-specific shrunken estimates across all years, the resulting model has an adjusted R^2 of 0.96. Thus, using the mean of the year-specific value-added estimates through the 2011–2012 school year is likely to allow me to correctly place teachers into the appropriate estimated value-added quintile (given that teachers are chosen from the first, third, and fifth value-added quintiles, creating substantial separation in value-added scores between the groups).

[47] The teacher identification numbers were provided by Mark Chin via email.

[48] For example, Appendix Figure D2 shows that the teacher with the highest value-added score in District B (i.e., the teacher with an estimated rank of nine) was predicted to have the lowest value-added score (i.e., the teacher has a predicted rank of one).

Appendix Figure D3 plots these means in order of increasing value-added within each district. We observe that there is generally a clear separation of teachers into low, medium, and high value-added groups based on these means and that each group contains three teachers, consistent with the design of the experiment.[49] Note, however, that distinguishing between teachers within a value-added quintile may be less reliable (e.g., all three teachers in the middle group in District 12 have similar mean value-added scores). Thus, my analysis treats the groups of three teachers based on estimated value-added, as opposed to individual teachers, as the unit of analysis.

Figure D2: Mapping Between Estimated Value-Added Rank and Predicted Value-Added Rank

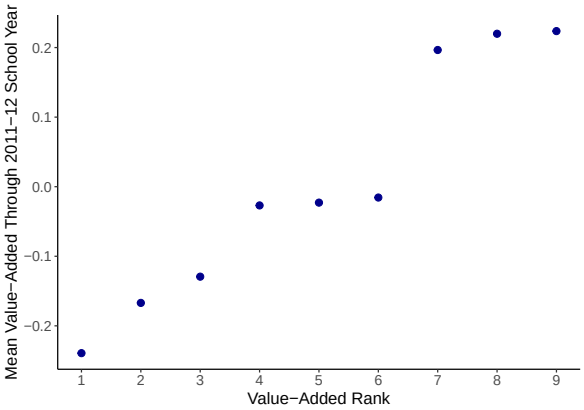


Note: This figure, which is Figure 2 of Hill et al. (2013), provides data on the value-added predictions made by the experts in the exercise described in Section 6. For each of the three districts involved in the analysis, the "Guessed VA Quintile" column shows the ordering of teachers based on predicted value-added, while the "Actual VA Quintile" column shows the ordering of the same teachers based on estimated value-added.

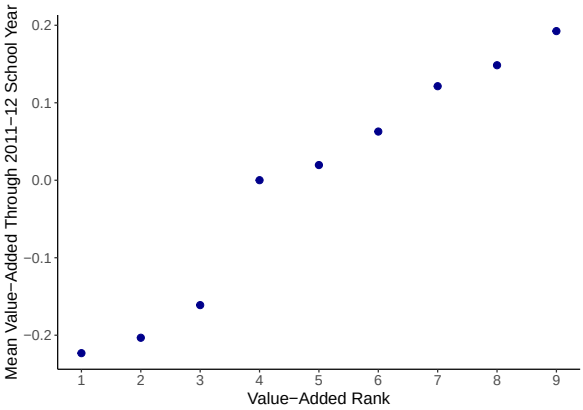
[49] Even in District 12, where the separation between the middle and high value-added groups is less visually clear than elsewhere, the difference in mean value-added scores between the sixth- and seventh-ranked teachers is still over a third of a standard deviation.

Figure D3: Mean Value-Added Through the 2011–2012 School Year

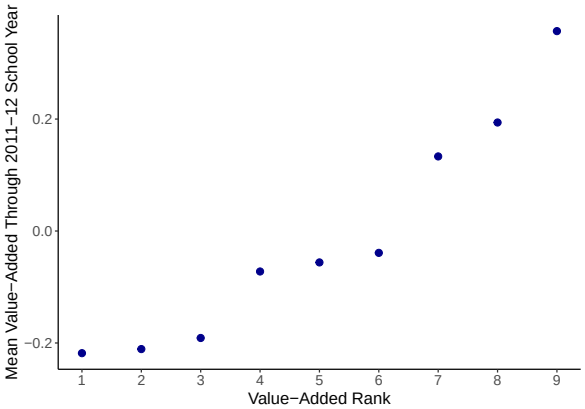
Panel A: District 11 (B)



Panel B: District 12 (R)



Panel C: District 14 (G)



Note: This figure plots the means of the year-specific shrunken value-added estimates through the 2011–2012 school year for the teachers included in the value-added prediction exercise described in Section 6. Within each district, observations are ordered by value-added rank, where a higher value-added rank corresponds to a higher value-added score.