Instructor Value-Added in Higher Education

Merrill Warnick*
Jacob Light†
Anthony Yim‡

October 16, 2024
Preliminary and Incomplete.
Please do not cite or circulate.
[Click here for latest version]

Abstract

Estimating instructors' value-added is challenging in post-secondary education because students can select their courses and instructors. In the absence of sound measures of value-added, universities use subjective student evaluations to make important decisions. We develop a method to estimate instructor value-added at any university which groups students together based on their past courses taken. We show that our non-experimental method controls for selection just as well as methods that exploit conditional random assignment of students to courses, using a unique policy change at a large public university in Indiana. We then apply our methods to 33 Texas public universities and standard forecast bias tests to demonstrate that our method controls for selection at a wide variety of post-secondary institutions. We find that individual instructors matter for students' future grades and post-college earnings in many subjects and courses. On average, moving to a 1 standard deviation better instructor would increase a student's next semester GPA by 0.13 points, and earnings six years after college entry by 0.17%. Strikingly, value-added is only weakly correlated with student evaluations, so an institution that made personnel decisions based on value-added rather than evaluations would improve student outcomes.

*Department of Economics, Stanford University. mwarnick@stanford.edu
†Hoover Institution. jdlight@stanford.edu
‡Department of Economics, Brigham Young University. anthony_yim@byu.edu

1 Introduction

Institutions of higher education strive to provide high-quality instruction, but often lack quantitative measures of instructor quality. Teaching is the focus of many universities, through which students develop the skills and knowledge to participate in a skilled workforce. Universities reveal their commitment to instructional quality by using it to make decisions about tenure-track promotions and retention. Commonly, institutions rely on subjective student evaluations to measure this quality. However, existing research suggests that these evaluations may distort incentives in the classroom[1] and reflect students' internal biases[2].

Quantitative measures of instructor quality, such as value-added models extensively applied in K-12 education, are an appealing alternative, but their application in higher education is complicated by substantial identification challenges. Value-added estimation requires an assumption of selection on observables: that the channels through which students can sort across instructors are observable and can be controlled for. In K-12, it is often sufficient to control for lagged student achievement (measured by a previous year's standardized test score) to satisfy this assumption (Kane and Staiger, 2008; Chetty et al., 2014a). In most higher education settings, however, researchers lack a similar standardized measure of student ability to summarize sorting. Furthermore, since college students are free to choose their course schedules, unobservable characteristics likely guide students' choices of instructors in ways that cannot be addressed by controlling for achievement alone. For example, students may have different intentions for taking Organic Chemistry: some may take it as a pre-requisite for medical school, while others may take it on the path towards becoming a chemist. Informed by these intentions, these students may select different courses, or even different instructors of the same course, if they perceive that the instructors may appeal differentially to their interests.

When students' unobservable intentions are correlated with both their choice of instructor and future outcomes (e.g., subsequent course selection or career outcomes), conventional value-added estimates that do not address this form of sorting will be biased. The small literature studying instructor value-added in higher education often relies on unique policies at specific institutions to identify instructor value-added in higher education, but these methods are not applicable to most courses and universities (Carrell and West, 2010; DeVlieger et al., 2018).[3]

[1]For example, that instructors will lower grading standards to boost student evaluations (Nelson and Lynch, 1984; Eiszler, 2002).

[2]Chisadza et al. (2019). These biases, typically based on fixed instructor characteristics, are distinct from statistical bias in value-added estimates. Throughout most of the paper, we use "bias" to refer to statistical bias of estimators.

[3]For example, in Carrell and West (2010), freshmen students at Air Force Academy are randomized

This paper proposes a general method for estimating instructor value-added at many universities by augmenting value-added estimates with students' "course histories." Motivated by the example of students who pursue the same major but have different unobserved "types" that steer them to different instructors, our approach aims to overcome bias from student sorting by identifying value-added from the differences in outcomes of students who have previously taken similar classes but, for some given class, have different instructors. For example, we might be able to distinguish students in the same Organic Chemistry course as "medical school types," who have previously taken human anatomy and biology, or "chemist types," who have previously taken calculus. We use group students with similar unobservable type using hierarchical clustering on their course histories. By limiting comparisons that identify value-added to only take place within these groups, we are able to control for some of the unobservable differences that might bias conventional value-added estimates.[4]

We begin by comparing value-added with course histories to two alternative strategies for addressing student sorting in value-added using data from Purdue University. At Purdue, we leverage a policy that assigned students to courses randomly, conditional on submitted preferences, to apply the literature's preferred sorting solution in higher education. This policy allows us to estimate value-added to next-semester GPA under conditional randomization for comparison with our course history value-added estimates which only use methods that could be applied at any university. We also estimate value-added using only lagged achievement to control for sorting, as is common in the K-12 value-added literature, to document that course histories or a policy that restricts student sorting are indeed necessary for identification. We find that course history and conditionally random value-added admit very similar rankings of instructors, with a correlation of 0.85 between the two within-subject rankings.

To document the necessity of including course history in value-added, we show that estimates of value-added with lagged achievement are forecast biased at Purdue. Forecast bias occurs when unobservably higher-ability students sort to higher-quality teachers. When estimates are forecast biased, value-added does not predict student outcomes 1:1 out of sample. We use two forecast bias tests to show that lagged achievement value-added estimates — which control for observable measures of student achievement, student characteristics and classmate characteristics — do not predict student outcomes out of sample.[5] Therefore,

into a set of required courses, removing the ability of students with different intentions to choose different instructors.

[4]Dale and Krueger (2002) provide a similar motivation for identifying the return to different universities by comparing outcomes of students who apply to the same set of universities but differ in the institution they ultimately attend. We extend their clever intuition to a setting where sorting across instructors, once we restrict comparisons to be between students with similar course histories, is more plausibly random.

[5]The first test uses changes in average jackknifed value-added to predict changes in average student

these controls are insufficient to control for student sorting.

Augmenting the value-added estimation with controls for students' course histories yields estimates of value-added that are substantially less biased than the estimates with only lagged achievement. The bias tests indicate that our value-added estimates with course histories predict student performance out of sample much better than lagged achievement value-added. We also find that value-added with course histories controls for sorting nearly as well as estimates that use the conditional random assignment policy.

The primary advantage of our course history method over using a policy to control for student sorting is that our method can be applied at any university. We apply our method to 33 four-year universities in Texas, where we observe detailed transcript-level data linked to post-graduation earnings. In Texas, we estimate value-added to both next-semester GPA and to log earnings six years post college entry. We apply the forecast bias test to both course history and lagged achievement value-added these Texas universities. The bias tests show that for both outcomes and nearly all universities, estimates of value-added that control for course histories are much less biased than estimates that control only for lagged achievement.

After confirming that value-added estimated with course history controls is forecast unbiased in Texas, we show that individual instructors affect both their students' next-semester GPA and earnings six years post college entry. An instructor with a 1 standard deviation higher value-added to GPA increases their students' next-semester GPA by 0.13 grade points on average. An instructor with 1 standard deviation higher value-added to earnings increases student earnings six years post college entry by 0.17% on average.

We investigate the characteristics of high value-added instructors. Although we document correlations between value-added and instructor characteristics, instructor characteristics are generally not strong predictors of value-added, consistent with similar findings in the K-12 literature.[6] On average, Black instructors have lower value-added to GPA compared to instructors of other races and non-native instructors have lower-value added compared to instructors born in the US. Instructors with lower academic rank have higher value-added to GPA, on average, although the differences across ranks are not statistically significant. For value-added to earnings, we find that female instructors have higher value-added compared to male instructors. Associate professors and non-tenure track instructors have significantly higher value-added, on average, compared to full professors and assistant professors.

outcomes within a course, similar to Chetty et al. (2014a). This test is identified by year-to-year changes in teaching rosters that are unexpected by students. The second test uses value-added estimated before the conditional random assignment policy to predict student outcomes during the policy period, controlling for random assignment.

[6]For instance, [NTD-cite] show that while X is the strongest predictor of teacher value-added, its predictive power is quite limited.

Finally, we investigate the relationship between subjective student evaluations and value-added measures of instructor quality using evaluations from three Texas universities. We find that student evaluations scores are mildly positively correlated with value-added to GPA, but are uncorrelated with value-added to earnings. Interestingly, the strongest predictor of student evaluations is not value-added, but instructor leniency. The correlation between evaluation scores and the average grades an instructor assigns is stronger than the correlation with either measure of value-added. Given that universities currently rely on student evaluations for personnel decisions, our findings suggest there may be welfare gains from incorporating value-added measures into assessments of instructor quality.

This project contributes to several strands of literature. A small but growing body of research estimates instructor value-added in higher education (Carrell and West (2010), DeVlieger et al. (2018), [NTD - add others]). These studies typically exploit unique institutional features, such as the random assignment of students to course sections or the use of standardized evaluations, to estimate instructor value-added for a limited set of courses. For example, Carrell and West (2010) leverages a unique policy of random assignment of students to core courses at the United States Air Force Academy to estimate value-added based on standardized final exam scores. A more recent study by DeVlieger et al. (2018) examines the value-added of algebra instructors at the University of Phoenix, a large for-profit online university. Both studies document large variation in instructors' impacts on student outcomes.

Our primary methodological contribution is to extend this approach to institutions that do not employ restrictive enrollment policies or standardized evaluations to control for sorting. Additionally, we estimate value-added for instructors across a broader range of courses than previous studies and extend the analysis to include value-added to earnings. Our results align with the existing literature in our finding of large variation in instructor value-added to student achievement. We additionally extend analysis by Carrell and West (2010) and DeVlieger et al. (2018) by comparing instructor value-added to student evaluations of instructor quality. In contrast to the finding in Carrell and West (2010) that instructors who raise student scores in their own courses tend to receive high evaluation scores but have low value-added to student grades in subsequent courses, we find that student evaluations are positively correlated with value-added to next-semester GPA but are uncorrelated with value-added to earnings.

We also build on a much larger literature that measures instructor value-added to test scores in K-12 education. This literature uses value-added methods to demonstrate that teachers in primary and secondary schools have causal impacts on student outcomes across

a variety of settings.[7] Our paper extends these methods to the higher education context and finds that university instructors similarly affect student outcomes. Ours builds most directly on three studies. Kane and Staiger (2008) estimate value-added in the Los Angeles Unified School District, using a randomized student-teacher assignment policy to validate estimates and test for bias. We apply a similar strategy, using a randomized sample to validate estimates from non-randomized data. Chetty et al. (2014a) develop forecast bias tests in New York schools, which we adapt for quasi-experimental tests based on semester-to-semester changes in teaching rosters. Chetty et al. (2014b) extend this work to estimate the effects of teachers on long-term outcomes like college attendance and earnings. We similarly estimate value-added to post-college wages and other long-term outcomes.

[NTD-Other literature to cite: Dale & Kruger, applications of ML to economics, inequality in higher ed, human capital vs signalling - point to impact (if any) of instructors, any other relevant metrics literature, student evaluations literature, lalonde's comparing observational and experimental stuff, Rockoff and Speroni, subjective and objective measures of instructor quality]

2 Statistical Framework and Estimation

Our statistical framework models the causal impact of an instructor on student outcomes as a fixed effect in a linear model. This framework motivates how we estimate individual instructor value-added measures using empirical Bayes shrinkage to account for measurement error. Our general model and estimation framework both follow the value-added literature.

We describe the value-added framework for a generic achievement measure A.[8] Following the value-added literature, we express the achievement A^*_{ijsct} of student i in course c of subject s during academic period t in instructor j's classroom as

$$A^*_{ijsct} = X_{it}\beta + C_{jsct}\gamma + \rho_c + \lambda_t + \nu_{ijsct} \tag{1}$$

$$\nu_{ijsct} = \mu_{js} + \epsilon_{ijsct} \tag{2}$$

[7] For example, Boardman and Murnane (1979); Hanushek (1979); Rockoff (2004); Jacob and Lefgren (2008); Rothstein (2010); Chetty et al. (2014b); Angrist et al. (2017); Macartney et al. (2018); Altonji and Mansfield (2018); Rose et al. (2022). We also add to a growing literature that estimates value-added to non-test outcomes, such as student behavior (Jackson, 2018; Petek and Pope, 2023) and academic performance far into the future (Gilraine and Pope, 2021).

[8] In elementary and secondary education, the conventional achievement measure is performance on a standardized test. Lacking standardized tests in higher education, we often use the students GPA in a future semester as our main achievement measure.

where X_{it} captures student i's background characteristics, C_{jsct} are characteristics of other students taking class c with instructor j,[9] ρ_c is a course fixed effect, λ_t is a period fixed effect and ν_{ijsct} is a composite error term, which contains individual error ϵ_{ijct} and μ_{js}, which is instructor j's value-added to $A_{ijs,t}$. Note that, in practice, all estimation will be taking place within-subject because there are few instructors who teach in more than one subject, so we will drop the s subscript for the remainder of the section.

Our estimands of interest are μ_j, instructor j's value-added, and the variance of the distribution of value-added, σ^2_μ, which describes the impact of moving to a higher-quality instructor. In particular, the standard deviation of the value-added distribution σ_μ is the average impact of having a one s.d. higher VA instructor.

We make two notable assumptions for our estimation. First, we assume that value-added is fixed across time t and across course c.[10] Next, we assume that both μ_j and ϵ_{ijct} are distributed normally, allowing us to use maximum likelihood estimation (MLE) to estimate variances for both distributions, following Gilraine et al. (2020).[11]

To estimate μ_j, we begin by residualizing A^*_{ijsct} on background and classroom characteristics:

$$A_{ijct} = A^*_{ijct} - \left(X_{it}\hat{\beta} + C_{jct}\hat{\gamma} + \hat{\rho}_c + \hat{\lambda}_t\right) \tag{3}$$

where the estimated coefficients and fixed effects come from the regression in equation (1), but including the instructor fixed effect μ_j. Including these fixed effects in the residualizing step assures that we estimate these coefficients using only within-instructor variation and not across-instructor variation (Chetty et al., 2014a).

Since value-added estimates are subject to classical measurement error due to the typically large variance in the student error term, ϵ_{ijct}, we shrink these estimates for use in bias tests. Consistent with the value-added literature, we apply empirical Bayes methods to shrink the estimates toward a normal distribution with a mean of zero. Intuitively, empirical Bayes down-weights the contribution of periods where an instructor has few students, and shrinks overall estimates for instructors when the estimated variance of the value-added distribution is smaller relative to the variance of the individual error distribution.[12]

[9] Specifically, we control for average lagged GPA of all students who take instructor j's sections of course c in period t. We define these characteristics at the instructor-course-period level, such that, for an instructor who teaches multiple sections of c in semester t, we pool across all of their sections.

[10] By defining value-added at the institution-by-subject level, value-added may vary for the very small set of instructors who are attached to multiple subjects or institutions.

[11] See Appendix B.1 for more details on this method.

[12] For additional detail, see Appendix B.2.

In some applications, such as the teaching roster changes bias test introduced in Section 4.2, we regress student characteristics on the estimated value-added of their instructors. If these students' data was used to estimate that value-added, there would be spurious correlation between estimated value-added and student outcomes. In order to avoid these spurious correlations, we construct jackknifed versions of empirical Bayes value-added estimates. Jackknife value-added estimates use all periods other than the current period to predict value-added in the current period, eliminating these correlations.[13]

When value-added is on the left hand side of a regression model, the classical measurement error in our value-added estimates does not bias the coefficients of the regressions, and we simply estimate value-added as the average of residuals A_{ijct}. This estimate is the same as if we added the instructor fixed effect to a regression based on equation (1). For each application, we will note which of the three versions of value-added is being used.

These measures of value-added are unbiased when $E[\epsilon_{ijct}|X_{it}, C_{jct}, \rho_c, \lambda_t] = 0$, or when students select instructors only on observables. Research suggests that this selection on observables assumption holds in the K-12 setting, but there is little evidence on when this assumption holds in the higher education setting (Kane and Staiger, 2008; Chetty et al., 2014a). We explain this assumption in greater detail in Section 4.

3 Data and Setting

We use two panels of administrative student data: transcript data from Purdue University, and linked transcript-to-earnings data covering all public universities in Texas. For both panels, we restrict our attention to undergraduate students.

3.1 Purdue Student Panel

We use data from Purdue University, a selective public institution in Indiana with a strong focus on STEM and Engineering and classified as an R1 university due to its high level of research activity. Our dataset covers student transcripts from 2011 to 2023, providing detailed information on course enrollment, grades, pre-enrollment characteristics (such as entrance exam scores), and student demographics.

Purdue instituted a policy starting in Fall 2018 that assigned students to courses via an algorithm based on ranked lists of course preferences. In Fall 2018 and Fall 2019, this process applied to first-time freshmen, and beginning in Fall 2020, nearly all students were

[13] Appendix B.3 contains more information about jackknife estimates.

assigned to courses through this algorithm each semester, including the Spring semesters[14]. In many cases, the algorithm randomly broke ties between students who had equal priority and similar preferences. Although students could influence which courses they were assigned to by ranking them, they had very little control over the instructor they would receive within each course. While students had the option to request specific sections or instructors, very few took advantage of this option. We observe the ranked lists of course preferences for students in every semester where students were assigned via algorithm: Fall 2018, Fall 2019, and every semester from Fall 2020 and onwards, including Spring semesters.

This random assignment policy allows us to construct benchmark estimates using methods similar to existing value-added estimates in the higher education literature for comparison with our course history value-added estimates. Additionally, it enables us to conduct a forecast bias test using the conditional random assignment.

3.2 Texas Student Panel

We use administrative data from the Texas Education Research Center, which contains linked transcript-to-earnings records for all students who attended public four-year universities in Texas from 2011 to 2021. The transcript data record every course taken by each student, including the instructor of record, as well as pre-enrollment characteristics (such as entrance exam scores) and student demographics. The Texas data are linked to the state's unemployment insurance system, allowing us to track quarterly earnings for students who remain employed in Texas.[15] Additionally, the Texas dataset includes information about instructors, such as their rank, demographic characteristics, and salary.

The Texas data offer two key advantages for our analysis. First, we can link student records to instructor characteristics and earnings, enabling us to estimate novel measures of instructor value-added to *earnings*, which is rarely possible in higher education, and explore the relationship between instructor characteristics and value-added. Second, the breadth of institutions in the sample is considerable: we analyze data from 33 public universities operating continuously from 2011 to 2021.[16] These universities are broadly representative of public universities across the US: on average, they admit 83% of applicants (compared to 78% for all US public universities), have similar student-faculty ratios (19.36 compared to 16.68 overall), and admit students with comparable standardized test scores. This diversity

[14]A few students, such as athletes, were exempt from the algorithmic assignment, and these students are excluded from our analysis.

[15]Earnings for self-employed workers or those who move out of Texas are not observed. We observe earnings for 85% of students, so this limitation likely has minimal impact on our analysis.

[16]Our analysis is limited to baccalaureate-granting institutions. We exclude the Texas A&M: San Antonio and the University of North Texas at Dallas, which do not provide consistent data over the full period.

makes the Texas data more reflective of typical US public universities than Purdue, which is relatively selective. While the Texas sample includes some highly selective institutions, such as the University of Texas at Austin and Texas A&M University, it also includes many non-selective institutions. Appendix Table 7 compares the characteristics of the Texas sample and Purdue to the typical US public university.

In Section 6.3, we compare our estimates of instructor value-added to the type of data many universities currently use to evaluate instructor quality: student evaluations. For this purpose, we collected instructor evaluations by scraping online archives from three Texas institutions: Texas Tech, Sam Houston State University, and the University of Texas at Tyler. These evaluations, dating back as early as 2006, include overall student impressions of the instructor/course as well as responses to specific questions, such as whether the course was conducted fairly or whether the instructor was approachable. We merge the evaluations data with the value-added data by matching course IDs across both datasets.

4 Lagged achievement is insufficient to control for bias from students' selection to instructors

In this section, we demonstrate that controlling for lagged measures of student achievement, which is typically sufficient in K-12 value-added estimation, is insufficient for eliminating bias from sorting in value-added estimation for higher education. We make this point at Purdue, where we leverage a unique institutional policy that randomly assigns students to courses and instructors conditional on their preferences to benchmark forecast bias of value-added with conventional lagged achievement controls to forecast bias of value-added under conditional randomization.

4.1 Value-added estimation using established methods

Research in K-12 value-added has shown that observable characteristics are often sufficient to control for student sorting across teachers (Kane and Staiger, 2008; Chetty et al., 2014a). The two most important among these characteristics are lagged student achievement (measured through prior standardized test scores) and class composition (measured through averages of these lagged scores). Standardized test scores correlate with unobservable ability that is likely correlated with both future performance and instructor assignment. Classroom averages of lagged test scores are a sufficient statistic for sorting patterns of other students in the course the instructor (Altonji and Mansfield, 2018).[17] These, along with a few other background

[17]For example, a high-achieving student may be assigned to a teacher that teaches "gifted and talented" students.

characteristics, control for sorting in part because much of the classroom assignment happens centrally.

In contrast, students at most universities have the freedom to choose both their courses and their instructors. When these choices correlate with unobserved student types that are correlated with future outcomes (e.g., motivation, work ethic, intended field of study, access to resources within the university), value-added estimates that do not account for this sorting will be biased. For example, consider two types of students taking organic chemistry: "medical school" types take organic chemistry as a pre-requisite for medical school, while "chemist" types take organic chemistry to develop foundational knowledge for their research. For exposition, the medical school type takes organic chemistry primarily to fulfill a requirement for a competitive graduate program and has an incentive to find the instructor who maximizes their likelihood of receiving a high grade. The chemist type, on the other hand, takes organic chemistry to develop valuable skills, and therefore has an incentive to find the instructor who provides them the most human capital.[18] Estimating instructor value-added to lifetime earnings without accounting for these different types will introduce bias because the estimates would be unable to disentangle differences in earnings attributable to the instructor from differences in earnings typical of doctors relative to chemists.

Measuring student ability in higher education is a second challenge relative to value-added estimation in K-12. Unlike K-12 students, college students typically do not take standardized tests at the end of their courses, which means we lack a consistent outcome measure of student achievement. Additionally, we do not have reliable measures of prior student achievement, which have been shown in K-12 research to be important for controlling for student sorting. While college entrance exams provide some pre-college measure of ability, they are not tied to specific courses. Furthermore, pre-college achievement may be less relevant as a lagged measure of student performance, particularly for older students. Similarly, GPA from previous semesters is insufficient because grades are a broad and imprecise indicator of student ability. Although these controls help account for student sorting, they may not fully eliminate bias.

To assess the extent of the challenges introduced by these limitations, we compare value-added estimates based solely on the conventional K-12 controls (lagged achievement and a set of fixed student characteristics) to estimates that utilize a unique institutional feature, which allows us to introduce additional controls such that student assignment to instructors for a given course is effectively random. This comparison will help determine the potential

[18]Students could be sorting to instructors in other ways as well. For example, all students might choose instructors with easier grading standards. If all students sorted in the same way for every course, forecast bias would not be an issue in higher education. The following section will demonstrate that forecast bias is a problem in higher education.

bias introduced when using only traditional controls versus a more robust set of controls that account for sorting more effectively.

For the approach that estimates value-added using only the controls used in K-12 estimation, we control for lagged student achievement using a student's semester GPA from the previous semester and incoming standardized test scores.[19] In addition, we control for a set of student and classroom characteristics: the student's level (freshman, sophomore, etc.), gender, race, and age, and classroom averages of lagged semester GPA and entrance exam scores. These controls represent the full set of characteristics used in the residualization step described in Equation 3.

We compare the estimates generated under the above procedure to estimates in a highly specific setting where students are assigned at random to instructors, conditional on their preferences. Such a setting is not typical in higher education, Purdue provides us a setting with such a policy. Purdue's course assignment policy operates fully through the ranked lists of preferences submitted by students. For example, students with identical preferences who want to enroll in an over-subscribed course are split randomly across sections using an algorithm. In this setting, we are able to control for student sorting by directly controlling for student preferences.

To control for student preferences, we augment Equation 1 with a preference similarity group fixed effect P_{ict}:

$$A^*_{ijsct} = X_{it}\beta + C_{jsct}\gamma + P_{ict} + \rho_c + \lambda_t + \mu_{js} + \epsilon_{ijsct} \tag{4}$$

The fixed effect P_{ict} partitions students in course c during period t into groups of students with similar lists of ranked preferences. Students with similar preference lists have the same or similar probability of being assigned to a given instructor, giving us randomization conditional on these preferences.[20]

4.2 Forecast bias in value-added estimates

To compare value-added estimates from lagged achievement models and conditional randomization approaches, we follow the existing literature by testing for forecast bias. Forecast bias arises when unobserved factors influencing the selection of students to instructors are

[19]Because summer academic periods are quite different from school-year academic periods, we use only data from fall and spring semesters to construct lagged and future GPA variables.

[20]We constructed these preference groups by applying hierarchical clustering to a vector containing indicators for the student's six most preferred courses. Appendix B.4 discusses hierarchical clustering in more detail.

correlated with student achievement, beyond what is captured by the controls in the model.

The intuition behind forecast bias is straightforward. When value-added estimates are biased due to student sorting, instructor effects tend to be overstated. This happens because high-ability students are more likely to be assigned to higher-quality instructors, making these instructors appear more effective than they truly are. Conversely, lower-quality instructors may seem worse because they are assigned lower-ability students. As a result, the variance of the estimated value-added distribution is distorted.[21]

Following Chetty et al. (2014a), we define the forecast bias B of value-added estimator $\widehat{\mu}_j$ as $1-\alpha$, where α is estimated from the regression of residualized achievement A_{ijct} on $\hat{\mu}_j$:

$$A_{ijct} = \rho_c + \lambda_t + \alpha\widehat{\mu}_j + \psi_{ijct} \tag{5}$$

If value-added estimates are not biased, α should equal 1. Intuitively, the forecast bias B captures the extent to which value-added measures fail to accurately predict residualized student achievement, often due to unobservable factors ψ_{ijct} influencing student-instructor assignments. When there is forecast bias, the true impact of an instructor who is one standard deviation above the mean is not simply $\sigma_{\hat{\mu}}$, the standard deviation of the estimated value-added distribution, but $(1-B)\sigma_{\hat{\mu}}$.

We apply two tests for forecast bias: one that uses changes in teaching rosters and one that uses Purdue's conditional random assignment policy.

The first forecast bias test, which we call the teaching roster changes forecast bias test, or "roster test" for short, followsChetty et al. (2014a) by using year-to-year variation in teaching assignments as a quasi-experimental source of variation. This variation arises from instructors shifting teaching responsibilities across semesters due to factors like sabbaticals, leaves, or changes in course load. By comparing changes in the average value-added of instructors within a subject and course level to corresponding changes in average student outcomes, we can assess whether the value-added estimates are forecast biased. If the value-added estimates are unbiased, a change in the average value-added for a course (e.g., replacing a low value-added instructor with a high value-added one) should predict a change in average student outcomes one-for-one. However, if the estimates are biased, these changes in value-added will not predict student outcomes as expected. This approach allows us to check whether the value-added estimates systematically overestimate or underestimate instructors'

[21]Note that students with high ability need not always choose instructors with high value-added to generate forecast bias. Forecast bias occurs and is detectable whenever students choose any instructor based on their unobservables, regardless of instructor characteristics or abilities. High ability students choosing high value-added instructors is the simplest example of this kind of sorting.

true impacts by using real-world shifts in instructor assignments as a natural experiment.

Formally, let $\overline{A_{slt}}$ represent the student-weighted average of residualized student achievement A_{icsjlt} within a subject-course level-period cell and $\overline{M_{slt}}$ represent the student-weighted average of jackknife empirical Bayes estimates of value-added within that same cell. Define the difference in average residualized achievement between periods[22] $\overline{A_{slt}}$ as:

$$\Delta A_{slt} = \overline{A_{slt}} - \bar{A}_{sl,t-2}$$

Define ΔM_{slt} analogously. The forecast bias test regresses changes in average residual student outcomes on changes in average value-added:

$$\Delta A_{slt} = \delta \Delta M_{slt} + \xi_{slt} \tag{6}$$

An estimate of $\widehat{\delta} = 1$ indicates that the estimates are forecast unbiased.

This test is identified by changes in teaching roster for a particular subject and level from semester to semester. These changes come from instructors taking leave of absence, going on sabbaticals, or simply moving to a different course load. We include subject fixed effects to ensure that we only use variation within a subject, since value-added estimation occurred within subject. By including additional fixed effects (e.g., subject-by-semester), we can ensure that the identifying variation comes from changes within the same semester and subject but across different course levels — for example, when an instructor switches from teaching a freshman-level course to teaching a senior-level course in the same subject.

This test is valid as long as changes in student unobservables ξ_{slt} are unrelated to these changes in teaching rosters: namely, if students do not time their enrollment to study with high value-added instructors. We test this assumption with a robustness check that regresses changes in student enrollment on changes in value-added, and find that changes in value-added within a course do not predict changes in student achievement. The full results of this test are in Appendix C.3.

Table 1 shows results of the roster forecast bias test for estimates of value-added to next-semester GPA at Purdue. We estimated value-added using lagged achievement during the random assignment period (2018-2023) for comparison with the conditional random assignment approach and over the entire data period (2011-2023). Columns 1 and 2 show results for these two estimates. The point estimates, 0.280 and 0.249 respectively, provide strong evidence that value-added based on lagged achievement is subject to forecast bias, indicating that lagged achievement alone does not adequately control for sorting.

[22]Note that we difference fall semesters with fall semesters and spring with spring because of the seasonality of courses.

Table 1. Teaching roster changes forecast bias test for Purdue

	Lagged Achievement		Preference Controls	Course Histories	
Period:	2018-23	2011-23	2018-23	2018-23	2011-23
	(1)	(2)	(3)	(4)	(5)
Δ average value-added	0.280	0.249	0.793	0.706	0.910
	(0.201)	(0.092)	(0.135)	(0.123)	(0.092)
N	988	5,867	798	810	5,785
Lagged Achievement	X	X	X	X	X
Preference Controls			X		
History Controls				X	X

Notes: The teaching roster changes forecast bias test leverages year-to-year variation in teaching assignments to assess whether changes in residual student achievement are predicted by shifts in instructor value-added, with estimates regressing students' residualized next-semester GPA on changes in average jackknifed value-added. Columns (1) and (2) control for lagged achievement; Column (3) adds course preference controls; Columns (4) and (5) incorporate course history controls. Columns (1), (3), and (4) restrict to the conditional random assignment period (2018-2023), while Columns (2) and (5) estimate on the full period (2011-2023). Observations are at the subject-course level-period level. Standard errors are clustered at the period-subject level. An estimate closer to 1 indicates better controls for sorting.

In contrast, value-added estimates derived from the conditional random assignment approach effectively control for sorting. Column 3 shows that the forecast bias test yields a point estimate of 0.793 for value-added under random assignment, suggesting that the policy sufficiently restricts student sorting to identify value-added. Although the confidence interval does not contain 1, a point estimate of near 0.8 aligns with expectations from the value-added literature.[23]

The second forecast bias test, which we call the conditional random assignment policy forecast bias test, or "policy test," uses Purdue's enrollment policy to control for sorting. Let $\widehat{\mu}_j^{pre}$ be value-added estimated on data from 2011-2017, before the conditional random assignment policy. This test regresses individual student outcomes during the conditional random assignment period on these out-of-sample empirical Bayes value-added estimates:

$$A_{ijct} = \rho_c + \lambda_t + \alpha \widehat{\mu}_j^{pre} + P_{ict} + \psi_{ijct} \tag{7}$$

where P_{ict} are preference similarity group fixed effects from Equation 4. This out-of-sample test is similar to tests from the value-added literature where researchers use value-added to

[23]For instance, Kane and Staiger (2008) found a forecast bias point estimate of roughly 0.8 when estimating value-added using explicit random assignment of elementary students to teachers.

Table 2. Conditional random assignment policy forecast bias test for Purdue

	Lagged Achievement (1)	Course Histories (2)
Value-added	0.332	0.720
	(0.117)	(0.069)
N	163,653	158,875
Lagged Achievement	X	X
History Controls		X

Notes: The conditional random assignment policy forecast bias test estimates the explanatory power of value-added estimated before Purdue's algorithmic assignment policy on post-policy changes in students' GPA. The estimates come from a regression of residual next-semester GPA for student-course pairs in the conditional random assignment policy period (2018-2023, where the residualization removes preference controls to give conditional random assignment), on empirical Bayes value-added estimated from 2011-2018. Column (1) estimates pre-policy value-added with only controls for lagged achievement, while Column (2) adds course history group fixed effect controls. Observations are at the student-course-instructor-period level. Standard errors are clustered at the period-subject level. An estimate closer to 1 indicates better controls for sorting.

predict the performance of students who move to a new school, arguing that movers are not able to sort as effectively to teachers. In our setting, we directly control for student sorting in the out-of-sample period using Purdue's policy. This gives us a test that relies on conditional random assignment rather than quasi-experimental variation in teaching rosters.

Column 1 of Table 2 shows the results of the policy forecast bias test for value-added with lagged achievement. The forecast bias coefficient has a point estimate of 0.332, which is again far from 1. This test's results confirm that lagged achievement alone is not sufficient to control for unobservable student sorting.

The results in this section confirm that conventional controls from the K-12 value-added literature are inadequate for addressing the student sorting that is prevalent in higher education. Without properly accounting for this sorting, value-added estimates using these established controls will be biased. At Purdue, our unique institutional setting — where students with identical preferences for oversubscribed courses were randomly assigned to sections — allows us to effectively control for this sorting. However, this course assignment

mechanism is highly unusual and not a feasible approach for most institutions. In the next section, we propose an alternative non-experimental method for estimating value-added that utilizes data commonly available to universities. We demonstrate that this approach reduces bias as effectively as the random assignment controls used at Purdue.

5 Augmenting Value-added with Course Histories

Unlike our previous application in Purdue, most universities do not have policies that restrict student sorting across instructors. We propose a method to control for student sorting on unobservables in the absence of such a policy: grouping students based on their "course histories," or the set of other courses that a student has taken. We demonstrate that controlling for course histories reduces bias to a degree comparable to conditional randomization at Purdue. Having confirmed that our method performs as well as methods accepted in the literature, we further demonstrate the performance of our course histories controls in estimating value-added to GPA and earnings for the 33 public universities in Texas in our sample.

5.1 Estimating Value-added with Course Histories

Student sorting across courses and instructors poses a challenge for value-added estimation only if the model fails to account for the factors driving this sorting. Returning to our example of the two types of students who take organic chemistry — medical school types and chemist types — if these groups sort to different instructors, value-added estimates that do not account for this sorting will be biased.

In this section, we propose that students reveal their types through the courses they have previously taken, and that we can use these past courses to restrict value-added comparisons within student types. We define the set of courses taken prior to, and contemporaneously with, a given course as a student's **course history**.[24] Medical school type and chemist type students may differ in various ways that affect their outcomes, such as career goals, motivation, research interests, and social networks. They are also likely to differ in the

[24]We include both contemporaneous and past courses to better classify students earlier in their academic careers. Since students typically enroll in courses before the semester begins, their choices are made before being influenced by any instructor that semester. This approach also allows us to estimate value-added for first-year students in their first semester, who would otherwise be viewed uniformly if we only considered previously-taken courses.

At many universities, students adjust their schedules during the first few weeks of a semester. If instructor impacts occur during this "shopping" period, our estimates could be biased. However, we expect instructor effects to emerge later in the semester. A robustness check, where course histories are based strictly on past courses, supports this expectation by finding X.

courses they select: chemistry students may enroll in prerequisites for advanced chemistry courses like calculus, while pre-med students may take courses like human anatomy or biology from the medical school core.

If these course selections distinguish different student types in organic chemistry, partitioning students into groups based on their course histories and controlling for group fixed effects could reduce bias in our value-added estimates. The logic of this approach — that students reveal unobservable similarities through their behavior — is well established in education economics. For example, Dale and Krueger (2002) use a similar framework by comparing the earnings of students who applied to the same sets of colleges to estimate the return to attending different universities.[25] While students attending different universities may differ in significant ways, they argue, those who applied to the same colleges are more likely to be similar, and their final choice among these colleges is more plausibly random. Here, we might think that the choice of instructor for the same course is a more plausibly random choice made by students who are revealed to be similar than universities in the same choice set, where differences may be more salient.

Formally, we define a course history for student i in period t as the set of courses that student i has chosen to enroll in during periods $t' \leq t$. To estimate value-added with course history controls, we augment Equation 1 with a course history similarity group fixed effect H_{ict}:

$$A^*_{ijsct} = X_{it}\beta + C_{jsct}\gamma + H_{ict} + \rho_c + \lambda_t + \mu_{js} + \epsilon_{ijsct} \tag{8}$$

H_{ict} partitions students taking course c during period t into groups based on the similarity of their course histories. We create these course history similarity groups using hierarchical clustering on the course histories of students enrolled in the same course, across instructors.[26] We encode course histories as indicator vectors of all possible courses.[27] After clustering, approximately 30% of students are grouped into singletons. For the main analysis, we exclude students in singleton groups. A robustness check in Appendix [NTD] shows that pooling all singleton students into a single reference group within each course and period has little effect on the results.

The hierarchical clustering with course histories approach frequently produces groups of students that align with intuition about how course histories reveal students' future inten-

[25]Recent work by Bonhomme et al. (2022) generalizes this method for grouping individuals with similar characteristics, using an approach analogous to the one we describe here.

[26]For additional details about our hierarchical clustering approach, see Appendix B.4.

[27]We do not partition the data further, such as by instructor or grade in previous courses, due to computational and data limitations. Incorporating these characteristics may be a direction for future research.

tions. For example, during one semester at Texas A&M in an Organic Chemistry course, hierarchical clustering using course histories produces a group of students who seem to be chemist type and a group with students who seem to be medical school type. Panel A of Appendix Table 9 lists the five courses most frequently taken previously by students in each of these groups. The first group's top past courses were Engineering Mathematics, Computational Engineering, Fundamentals of Chemistry, and two courses from the Biomedical Engineering Degree, indicating that these students are likely chemist type students intending to pursue a degree in biomedical engineering. In contrast, the second group's most popular courses were from the Health core, indicating that these students were likely in Organic Chemistry to fulfill a medical school requirement. Tellingly, the majority of these students had also taken a psychology and sociology course that are strongly recommended for MCAT preparation, indicating that these students are likely medical school type students from our example.

We give two other illustrative examples of hierarchical clustering with course history's effectiveness at grouping students with similar future intentions. Panel B of Appendix Table 9 lists the top courses for students in two course history similarity groups in the same semester of Intermediate Microeconomics, again at Texas A&M. The first group of students had most frequently taken business courses, like management, accounting and marketing, and likely intend to pursue a career in business. The second group of students had most frequently taken courses in the Agricultural Economics program, and likely intend to pursue a career in agricultural economics. Panel C of lists the top courses for students in two course history similarity groups in Calculus 3 at Purdue. The first group of students had most frequently taken computer science core courses, and are likely "computer science" type students. The second group of students had most frequently taken engineering core courses, and many of the students had taken the course Computer Science with applications to Engineering. These students are likely "engineering" type students.

Differences in intentions only bias value-added when they are correlated with how students sort to instructors. We test if students in our course history groups are sorting systematically to certain instructors by conducting Pearson's χ^2 independence tests. The final row in each panel shows p-values from these tests. In each case, we reject the null hypothesis that course history similarity groups and instructor choices are independent at the 10% confidence level. In the case of Organic Chemistry and Calculus 3, we reject the null at the 5% confidence level.

This differential sorting to instructors across groups of students with similar course histories could not have been controlled for by lagged achievement alone. For example, Panel B shows average last-semester GPA and entrance exam percentile separately for the busi-

ness type students and the agricultural economics type students. The mean lagged GPA for these groups is nearly identical and mean entrance exam scores are very similar across groups. Furthermore, figure [NTD] shows substantial overlap between the distributions of lagged GPA and entrance exam scores for these two groups of students. In calculus 3 at Purdue, computer science and engineering type students also have substantial overlap in both prior achievement measures, as shown in Appendix Figure 3. This means that controlling for only lagged achievement and not course history groups would have used cross-history group comparisons to identify value-added. These comparisons likely result in biased estimates of value-added for instructors of both courses since these students seem to differ in their future intentions.[28]

Course histories addresses the sorting challenge in these three examples, but whether they address sorting more generally is an empirical question. In the following section, we conduct forecast bias tests to show that course histories do indeed control for student sorting.

5.2 Forecast bias estimates for course history value-added at Purdue

To test whether controlling for course histories effectively addresses bias from students' course and instructor selection, we apply both forecast bias tests described in Section 4.2 to value-added estimates that incorporate course history controls. Table 1 compares results from the roster forecast bias test with results using the lagged achievement and conditional random assignment approaches. We estimate value-added using course histories for both the full Purdue panel and the period following the implementation of Purdue's conditional random course assignment policy.

The forecast bias estimates for value-added using course history controls show a large improvement over estimates that control only for lagged achievement. On the full Purdue panel, the forecast bias coefficient is 0.910. When estimated solely during the post-randomization period, the coefficient is 0.706, which is relatively close to the estimate obtained from the conditional random assignment approach. We suspect that the shorter panel, which overlaps with the Covid-19 pandemic, complicates value-added estimation during the post-randomization period.

We also conduct the policy forecast bias test, which estimates value-added using data from pre-2018 and predicts student outcomes from 2018-2023, controlling for student sorting using the conditional random assignment policy. Table 2 compares the forecast bias coefficient estimates for value-added estimated using lagged achievement controls and course histories.

[28]It is possible that the failure of lagged achievement is in part due to the fact that the widely available measures of student achievement in post-secondary education, GPA and SAT scores are not perfect measures of student ability in the course. However, no better measure of applicable student achievement is readily available at the majority of institutions.

Column 2 shows the results for course history value-added. With a point estimate of 0.72, course histories address unobservable student sorting much better than lagged achievement alone.

Finally, we demonstrate that our instructor value-added estimates with controls for course histories align with instructor value-added estimates that leverage conditional random assignment. Since we estimate course history and conditional random assignment value-added using only students in non-singleton groups, each measure estimates value-added for different groups of students. [NTD-table with the tab of ungrouped] This sample difference means that these two measures of value-added are somewhat different. To make the estimation samples more comparable, we restrict the sample to the set of students who were in non-singleton groups and in non-singleton course history groups. Even with this sample restriction, the actual estimation samples will still differ slightly, as some non-singleton history groups lose some of their members and become singleton groups because they were ungrouped for preferences, and vice versa. In part due to some of these sample differences, the raw correlation between course history and conditional random assignment value-added under the restrictions is only 0.54. Even so, the correlation between course history value-added and conditional random assignment value-added is much larger than the correlation between lagged-achievement and conditional random assignment value-added, which is only 0.28.

Though course history and conditional random assignment value-added are only moderately correlated, we find that they admit very similar rankings of instructors. The correlation between within-subject instructor ranks of value-added with course histories and within-subject instructor ranks value-added with conditional randomization is 0.83. Rank comparisons between value-added estimates are an important exercise since value-added is always a relative measure: quality is measured relative to the mean, within a subject. Since course history and conditionally random value-added make very similar distinctions between low- and high-value-added instructors, course history value-added is indeed addressing student sorting at Purdue.

5.3 Forecast bias estimates for course history value-added in Texas

The forecast bias tests in the previous section confirm that value-added estimates that control for course histories substantially reduce bias in value-added estimation from student sorting to instructors. One highly appealing feature of this method is that it can be applied at any university. For the rest of the paper, we focus our attention on value-added estimation in Texas, where we have linked transcript-to-earnings data for all 33 public universities.

In Texas, we estimate value-added for both next-semester GPA (as at Purdue) and future

earnings. The ability to estimate value-added to earnings is a unique feature of the Texas dataset. Specifically, we estimate value-added to (log) earnings six years after a student enters college.[29]

We first verify that these value-added measures are forecast unbiased. To do so, we apply the teaching roster change forecast bias test, using quasi-experimental variation in teaching rosters to identify the forecast bias coefficients. Figure 1 plots the results from each university for the forecast bias tests, estimated with lagged controls only (the blue dots) and with controls for course histories (the red dots).

The results of the forecast bias tests indicate that using course history controls effectively accounts for unobservable student sorting. As was the case in Purdue, value-added estimates that control only for lagged achievement suffer from substantial forecast bias. For value-added estimates of next-semester GPA, incorporating course histories significantly reduces forecast bias compared to using only lagged achievement. Most estimates fall between $0.8 - 1.2$, with many even closer, resulting in a median forecast bias estimate of 0.17. Similarly, course histories help reduce forecast bias in value-added estimates of earnings. For all but one university, course history controls reduce bias, with a median forecast bias of 0.15 for value-added to earnings.

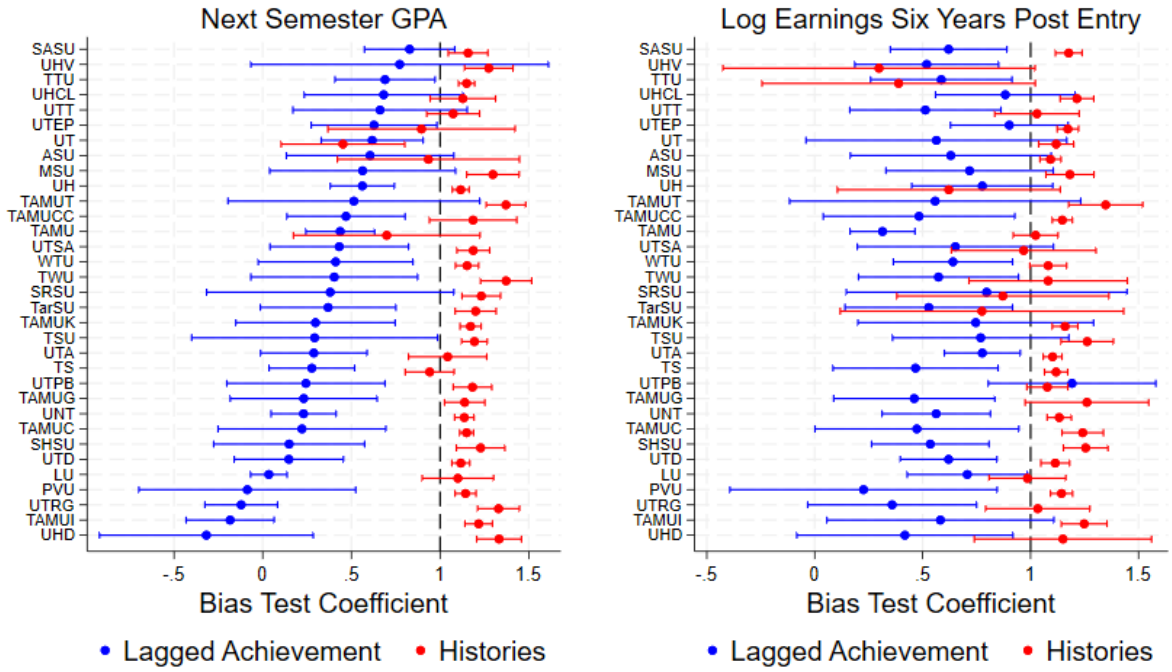
These findings demonstrate that course history controls are effective at controlling for student sorting across a wide range of universities. Having addressed the bias in value-added estimates, we next use estimates of instructor value-added to describe the characteristics of highly effective instructors.

6 Impacts and Characteristics of High Value-added Instructors

Having developed a method for estimating instructor value-added in higher education, this section uses our new method to assess whether instructors vary in their impacts on students' outcomes. The results of this section reveal that instructors do, in fact, impact students' future GPA and earnings. We then document characteristics of high- versus low-value-added instructors and compare our value-added estimates to students' subjective evaluations of instructor quality.

[29]Choosing the right year to measure earnings is challenging because students leave college, begin their careers, and reach the steady state of their earnings at different times. Ideally, we would measure value-added to earnings far enough into the future to avoid these timing issues. However, due to the relatively short panel of earnings data, we select the six-year mark, as this is the earliest point where most students are on a stable earnings trajectory and their earnings are high enough to exclude part-time jobs held during college. Our results remain consistent when using alternative measures of earnings.

Figure 1. Teaching roster changes forecast bias test for Texas universities



Notes: The teaching roster changes forecast bias test leverages year-to-year variation in teaching assignments to assess whether changes in residual student achievement are predicted by shifts in instructor value-added, with estimates regressing students' residualized next-semester GPA on changes in average jackknifed value-added. Bias coefficients estimated separately for each Texas university, controlling for period-subject fixed effects. Observations are at the subject-course level-period level. Standard errors are clustered at the period-subject level. An estimate closer to 1 indicates better controls for sorting.

6.1 Instructors impact students' future outcomes

Table 3. Variances of value-added distributions

	Next-term GPA (1)	Log Earnings (2)
All	0.018	0.028
R1	0.020	0.033
Non-R1	0.016	0.022

Notes: Variance of the value-added distributions were estimated within subject and institution, using maximum likelihood estimation, following Gilraine et al. (2020). This table then shows student-course-period weighted averages of these variances across subject and institution. Universities are split according to their Carnegie classification [NTD-pick the year to split]: R1 universities have "very high research activity."

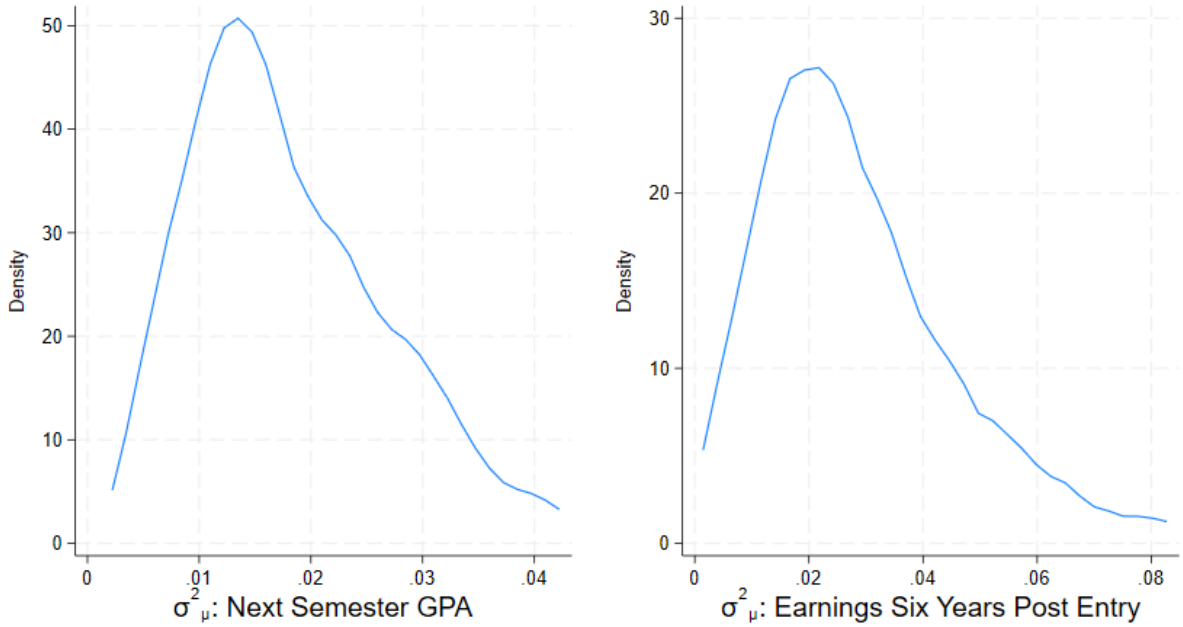
Table 3 summarizes the estimated variances in value-added to both next-semester GPA and future earnings, based on data from the Texas universities sample. These variances are substantial, indicating that instructors have meaningful impacts on students' short- and long-term outcomes. Taking the square root of these variances provides estimates of the effect of having an instructor with 1 standard deviation higher value-added than the average. We find that an instructor with 1 s.d. higher value-added to GPA increases their students' next-semester GPA by 0.13, which is approximately [NTD-recheck mean]% of the mean. This GPA increase is comparable to raising a student's average grade from a B- to nearly a B. While this variance is somewhat larger than the value-added variances observed for K-12 teachers in standardized test scores, it aligns closely with estimates of value-added for college algebra instructors on end-of-term exams DeVlieger et al. (2018). Instructors could influence students' future grades through multiple channels. For example, they may vary in their ability to enhance students' human capital, or they could affect students' future course selections.

Instructors also affect students' future earnings. An instructor with 1 s.d. higher value-added to earnings increases students' earnings by 0.17% six years after they enter college.[30]

[30]This estimate includes only students who have non-zero earnings in at least one quarter during the sixth year post-matriculation. We observe earnings for more than 80% of the students in our sample. Students without earnings data likely fall into two categories: those with no wage earnings during the year, and those working outside Texas (since Texas UI data only capture in-state earnings). We speculate that the latter group is larger. If out-of-state students have systematically different earnings than those working in Texas, and if instructor choice is correlated with students' likelihood of moving out of state, omitting these earnings may introduce bias into our value-added estimates.

Similar to value-added for GPA, instructors can influence future earnings through several pathways, such as increasing human capital, improving graduation rates, or altering students' field of study.

Figure 2. Distribution of subject-level variance in course history value-added



Notes: Smoothed kernel density plot of average variances of value-added across subjects within an institution. Variance of the value-added distributions were estimated within subject and institution, using maximum likelihood estimation, following Gilraine et al. (2020).

Instructor value-added varies across institutions and subjects. Figure 2 shows smoothed density plots of average value-added to future GPA and earnings within institution. Average value-added to GPA is relatively compressed, with variances between 0.003 and 0.17, corresponding to grade increases of 0.05 to 0.16 for moving to a 1 s.d. better instructor. Average value-added to earnings is more spread, with variances between 0.005 and 0.055, corresponding to income increases of 7% to 23% for moving to a 1 s.d. better instructor.

[NTD-investigate R1 split and look at skew of GPA and earnings] Rows 2 and 3 of Table 3 show average variance of value-added to next-semester GPA and earnings separately for the R1 (high research intensity) and non-R1 universities, respectively. On average, the variance of value-added to both GPA and earnings is higher at R1 universities than at non-R1

universities. The larger variance of value-added to earnings at R1 universities could be due to the fact that the raw variance of log earnings six years post entry is somewhat higher at R1s than non-R1s, with variances of 0.75 and 0.72 respectively. Since students at higher-ranked universities have higher "upside" jobs, especially immediately post graduation, instructors may have more ability to impact the early career earnings of their students.

Table 4. Variances of value-added for selected subjects

	Next-term GPA (1)	Log Earnings (2)
All	0.018	0.028
Education	0.020	0.014
Social Sciences	0.015	0.021
Computer Science	0.026	0.046
Engineering	0.024	0.053
Biology	0.016	0.036

Notes: Variance of the value-added distributions were estimated within subject and institution, using maximum likelihood estimation, following Gilraine et al. (2020). This table shows student-course-period weighted averages of these variances, across institution and within 2-digit CIP codes.

We also find that instructor impacts vary across subjects. Table 4 summarizes the average variance in instructor value-added across institutions, aggregated by two-digit CIP code. The first column of the table summarizes the variance in value-added to next-semester GPA. Fields with large fractions of in-major students taking courses such as Computer Science and Engineering, have higher variance of value-added to GPA. In contrast, fields like Biology and Social Sciences, which offer more courses to non-major students, show smaller variances in instructor value-added to GPA. One possible explanation for these differences is the likelihood that students will continue taking courses in these fields, which may influence the variance in value-added. Indeed, we find that the correlation between mean average value-added to next-semester GPA and mean persistence in the subject is positive (0.32).

The second column of Table 4 reports the variance in value-added to future earnings. The smallest variances in earnings value-added are in fields like Education, and the Social Sciences, while the largest are in Computer Science and Engineering. One interpretation of these findings is that instructors have more influence on future earnings in fields that lead to careers with greater income variability. For example, earnings for graduates in Engineering and Computer Science can differ significantly based on job placement, and the higher variance in value-added suggests that instructors may play a critical role in equipping students with

the skills needed to secure high-paying jobs. In contrast, graduates in Education, who often become teachers, typically have earnings determined by standardized pay scales, which show little variation within the same cohort. Indeed, we find that a subject's variance in earnings and the variance in value-added for instructors in that subject are positively correlated (0.29). These results also highlight the role of major choice as a central factor mediating the impact of instructor value-added on future earnings.

6.2 Characteristics of High-Quality Instructors

Understanding the characteristics of high-quality instructors is critical for universities as they seek to optimize resource allocation and make informed personnel decisions, such as how to adjust pay based on experience. In this section, we examine differences in instructor quality across various demographic and professional categories.

In Texas, we link instructor characteristics (e.g., rank, race/ethnicity, gender, age, salary) to our value-added measures.[31] We regress each value-added measure on these characteristics separately, controlling for institution- and subject-fixed effects, to evaluate how instructor characteristics correlate with instructor quality. Note that since value-added is estimated and normalized within institution and subject, these fixed effects are *not* picking up differences in the allocation of instructors with high value-added across subjects.

Table 5 summarizes the regression estimates. The coefficients represent the difference in average value-added associated with a given characteristic, relative to the omitted category (respectively: assistant professors, white, male, and native-born). For example, the estimate in the first row of column 1 in panel A suggests that full professors, on average, have value-added to next-semester GPA that is -0.005 GPA points lower than assistant professors. Additionally, for rank and race regressions, in panel A and B respectively, we report p-values for an F-test of coefficient equality as there are multiple categories. It is important to note that for some instructors who began teaching in later years, we cannot estimate value-added to earnings, as our estimation requires us to observe earnings up to six years after a student takes a given course.

On average, Black instructors and non-US citizens exhibit lower value-added to next-semester GPA compared to instructors of other races or nationalities. These correlations are both significant. Instructors of higher rank tend to have lower value-added to GPA than lecturers and instructors of lower rank. However, these correlation are not statistically significant, and an F-test of coefficient equality for instructors does not reject.

The estimates in Table 5 also suggest that the characteristics of instructors with high

[31]Since our primary value-added estimates are fixed across time, we take the modal observed academic rank.

Table 5. Heterogeneity in value-added across instructors

	VA to GPA (1)	VA to Earnings (2)
A: Academic Rank		
Full Professor	-0.005	0.000
	(0.004)	(0.004)
Associate Professor	-0.005	0.013**
	(0.004)	(0.005)
Non-Tenure Track	-0.001	0.009**
	(0.003)	(0.003)
N	69,406	62,722
F-test	0.264	0.009
B: Race		
Asian	-0.005	-0.004
	(0.003)	(0.004)
Black	-0.009**	-0.001
	(0.004)	(0.005)
Hispanic	0.002	0.002
	(0.003)	(0.004)
N	69,468	62,783
F-test	0.020	0.473
C: Female	0.000	0.004*
	(0.003)	(0.002)
N	69,468	62,783
D: International	-0.010**	0.000
	(0.003)	(0.004)
N	69,468	62,783

Notes: Estimates are from separate regressions of value-added on instructor characteristics, where each column in each panel estimates a separate regression. The omitted instructor rank category in Panel A is Assistant Professor; the omitted race in Panel B is white. Regressions control for subject and institution, with observations at the instructor level. Standard errors are clustered at the subject-institution level.

value-added to earnings may differ from those with high value-added to GPA. Specifically, we find that female instructors have significantly higher value-added to earnings than male instructors with a coefficient of 0.004. There is no strong evidence of a significant relationship between instructor race or nationality and value-added to earnings. For academic rank, we

find significant differences. In particular, we find that on average, full professors have the same value-added as assistant professors, while associate professors and non-tenure track instructors have significantly higher value-added than assistant professors. Furthermore, an F-test rejects the null hypothesis that these coefficients are the same.

Though we do find some significant correlations between instructor value-added and instructor characteristics, instructor characteristics do little to explain value-added. Table **??** [NTD-add table when R2 off the server] shows regressions of value-added on the full set of instructor characteristics shown in Table 5, and additionally includes age, average log salary, and fraction of students from upper-level courses taught by the instructors. The R^2 in both regressions is extremely small, with values of [NTD]. This result lines up with the value-added literature in K-12, which finds that observable characteristics do little to predict teacher value-added [NTD-find cites, i think there's a jacob and lefgren paper].

6.3 Correlation Between Value-added and Student Evaluations

In the absence of quantitative measures of instructor quality, many universities rely on student evaluations to assess teaching effectiveness. In this section, we investigate whether these subjective evaluations align with instructor quality as measured by value-added. To conduct this analysis, we scraped teaching evaluations for all courses from three institutions: Texas Tech, Sam Houston State, and the University of Texas at Tyler. At Texas Tech, we observe student evaluations from 2006-2023, at Sam Houston State, we observe student evaluations from 2005-2023, and at the University of Texas at Tyler, we observe student evaluations from 2018-2023. We calculated average evaluation scores for each instructor over these periods and merged these averages with the corresponding value-added estimates.

The student evaluation surveys vary across years and institutions. To create a standardized measure, we focused on a few categories of questions that appeared in most years and in all three institutions. Questions in the "Instructor Quality" and "Course Quality" categories capture students' general impressions of the instructor and the course. Questions in the "Teaching Quality" category specifically assess the instructor's effectiveness in teaching and conveying material. We standardized these average evaluation scores within each subject and institution to ensure comparability across time and courses.[32]

The first two columns of Table 6 summarize bivariate regression estimates of the relationship between instructor value-added to GPA and earnings on these aggregated evaluation scores. Instructor value-added scores in each column are standardized within subject and institution to be on a common scale. The first column shows correlations between value-added to GPA and evaluation scores are significantly correlated. The largest coefficient indicates

[32]For the exact questions, see Appendix [NTD].

that on average, instructors with 1 s.d. better Course Quality evaluations have 0.108 s.d. higher value-added to GPA. In contrast, we find that instructor value-added to earnings is not significantly correlated with any student evaluation score, again suggesting that there are within-instructor differences between value-added to GPA and value-added to earnings.

Motivated by previous work suggesting that student evaluations distort grading incentives in the classroom (Nelson and Lynch, 1984; Eiszler, 2002), we also investigate whether instructors who assign higher grades to students receive higher evaluation scores. In the third column of Table 6, we assess the relationship between student evaluations and instructor leniency to explore whether students prefer easier courses. We measure leniency by calculating the difference between the average grades assigned by an instructor in a course and the average grades given by other instructors teaching similar courses. The measures of leniency are then standardized within subject and institution for comparability of estimates.

We find that student evaluations are significantly correlated with leniency. Instructors who are more lenient tend to receive higher ratings, particularly in the "Course Quality" category, and also score higher in overall evaluations: an instructor with 1 s.d. higher course quality scores has 0.251 s.d. higher leniency scores. Furthermore, the correlation relationship between leniency and evaluations is stronger than the relationship between value-added to GPA and evaluations. For each category of evaluation questions, the correlation between leniency and evaluation scores is more than twice the magnitude than the correlation with value-added and evaluation scores. This provides suggestive evidence that students favor instructors who assign higher grades.[33]

7 Policy Implications

7.1 Student evaluations and non-tenure-track instructor retention

To understand how institutions actually use student evaluations in personnel decisions, we investigate the relationship between student evaluations an instructor receives and their probability of being hired at the same institution the next year.

7.2 Manipulating actual personnel policy

[in this subsection, we change the coefficients from the OLS model and see what would happen to retention and then figure out what would happen to student outcomes]

[33]Note that in this exercise, we do not control for backgrounds of students when constructing leniency measures.

Table 6. Comparison of instructor value-added to student evaluations

	Value-added		
	GPA	Earnings	Leniency
	(1)	(2)	(3)
A: Instructor Quality	0.101***	-0.019	0.218***
	(0.019)	(0.020)	(0.041)
N	5,668	5,257	6,116
B: Teaching Quality	0.099***	-0.028	0.221***
	(0.023)	(0.019)	(0.036)
N	5,668	5,257	6,116
C: Course Quality	0.108***	-0.019	0.251***
	(0.018)	(0.023)	(0.042)
N	5,055	4,697	5,472

Notes: Estimates are from separate bivariate regressions of instructor and course quality measures, based on student evaluations, on instructor value-added or leniency. Evaluations are from Sam Houston State University, Texas Tech, and the University of Texas at Tyler. The specific questions related to instructor, teaching, and course quality are detailed in Appendix [NTD]. Leniency is defined as the difference between an instructor's average grades and the average grades given by other instructors teaching the same subject at the same level. All student evaluation scores, value-added, and leniency measures are normalized. Regressions control for institution, with observations at the instructor level. Standard errors are clustered at the institution level.

7.3 Instructor de-selection

In the spirit of Hanushek (2009) and Chetty et al. (2014b), we evaluate the effects of implementing a policy that replaces the bottom 5% of non-tenure-track instructors with mean instructors.

8 Conclusion

References

Joseph G. Altonji and Richard K. Mansfield. Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects. *American Economic Review*, 108(10):2902–46, October 2018. doi: 10.1257/aer.20141708. URL https://www.aeaweb.org/articles?id=10.1257/aer.20141708.

Joshua D. Angrist, Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(2):pp. 871–919, 2017. ISSN 00335533, 15314650. URL https://www.jstor.org/stable/26495151.

Natalie Bau and Jishnu Das. Teacher value added in a low-income country. *American Economic Journal: Economic Policy*, 12(1):62–96, February 2020. doi: 10.1257/pol.20170243. URL https://www.aeaweb.org/articles?id=10.1257/pol.20170243.

Anthony E. Boardman and Richard J. Murnane. Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education*, 52(2):113–121, 1979. ISSN 00380407, 19398573. URL http://www.jstor.org/stable/2112449.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643, 2022. doi: https://doi.org/10.3982/ECTA15238. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15238.

Scott E. Carrell and James E. West. Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432, 2010. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/10.1086/653808.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9): 2593–2632, September 2014a. doi: 10.1257/aer.104.9.2593. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2593.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104 (9):2633–79, September 2014b. doi: 10.1257/aer.104.9.2633. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2633.

Carolyn Chisadza, Nicky Nicholls, and Eleni Yitbarek. Race and gender biases in student evaluations of teachers. *Economics Letters*, 179:66–71, 2019.

Stacy Berg Dale and Alan B. Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 117(4):1491–1527, 2002. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/4132484.

Pieter DeVlieger, Brian Jacob, and Kevin Stange. Measuring Instructor Effectiveness in Higher Education. In *Productivity in Higher Education*, NBER Chapters, pages 209–258. National Bureau of Economic Research, Inc, June 2018. URL https://ideas.repec.org/h/nbr/nberch/13880.html.

Charles F. Eiszler. College students' evaluations of teaching and grade inflation. *Research in Higher Education*, 43(4):483–501, aug 2002. ISSN 1573-188X. doi: 10.1023/A:1015579817194. URL https://doi.org/10.1023/A:1015579817194.

Michael Gilraine and Nolan G Pope. Making teaching last: Long-run value-added. Working Paper 29555, National Bureau of Economic Research, December 2021. URL http://www.nber.org/papers/w29555.

Michael Gilraine, Jiaying Gu, and Robert McMillan. A new method for estimating teacher value-added. Working Paper 27094, National Bureau of Economic Research, May 2020. URL http://www.nber.org/papers/w27094.

Eric A. Hanushek. Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources*, 14(3):351–388, 1979. ISSN 0022166X. URL http://www.jstor.org/stable/145575.

Eric A. Hanushek. Teacher deselection. In Dan Goldhaber and Jane Hannaway, editors, *Creating a New Teaching Profession*, pages 165–180. Urban Institute Press, Washington, DC, 2009.

C. Kirabo Jackson. What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107, 2018. doi: 10.1086/699018. URL https://doi.org/10.1086/699018.

Brian A. Jacob and Lars Lefgren. Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1):101–136, 2008. ISSN 0734306X, 15375307. URL http://www.jstor.org/stable/10.1086/522974.

Thomas J Kane and Douglas O Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Working Paper 14607, National Bureau of Economic Research, December 2008. URL http://www.nber.org/papers/w14607.

Hugh Macartney, Robert McMillan, and Uros Petronijevic. Teacher value-added and economic agency. NBER Working Paper w24747, National Bureau of Economic Research, 2018. URL https://ssrn.com/abstract=3202050. Available at SSRN: https://ssrn.com/abstract=3202050.

Jon P. Nelson and Kathleen A. Lynch. Grade inflation, real income, simultaneity, and teaching evaluations. *The Journal of Economic Education*, 15(1):21–37, 1984. doi: 10.1080/00220485.1984.10845044. URL https://www.tandfonline.com/doi/abs/10.1080/00220485.1984.10845044.

Nathan Petek and Nolan G. Pope. The multidimensional impact of teachers on students. *Journal of Political Economy*, 131(4):1057–1107, 2023. doi: 10.1086/722227. URL https://doi.org/10.1086/722227.

Jonah E. Rockoff. The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2):247–252, 2004. ISSN 00028282. URL http://www.jstor.org/stable/3592891.

Evan K Rose, Jonathan Schellenberg, and Yotam Shem-Tov. The effects of teacher quality on adult criminal justice contact. Working Paper 29555, National Bureau of Economic Research, July 2022. URL http://www.nber.org/papers/w29555.

Jesse Rothstein. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*. *The Quarterly Journal of Economics*, 125(1):175–214, 02 2010. ISSN 0033-5533. doi: 10.1162/qjec.2010.125.1.175. URL https://doi.org/10.1162/qjec.2010.125.1.175.

	Population		Texas Sample		Purdue	
	mean	sd	mean	sd	mean	sd
Enrollment	11,639	11,823	18,107	15,327	40,555	
Admit rate	0.78	0.18	0.83	0.15	0.53	
SAT-equivalent: 25 pctile	954	144	975	130	1,180	
SAT-equivalent: 75 pctile	1,182	127	1,167	125	1,410	
Average tuition	19,338	8,724	19,509	7,258	28,520	
Average price	13,766	4,172	11,774	2,999	11,898	
Student-faculty ratio	16.68	4.54	19.36	3.34	13	
6-year graduation rate	0.52	0.16	0.44	0.16	0.81	
Has a doctoral program	0.41	0.49	0.48	0.51	1.00	
R1 (very high research intensity)	0.12	0.33	0.09	0.29	1.00	
R2 (high research intensity)	0.12	0.33	0.18	0.39		
Other Carnegie classification	0.75	0.43	0.73	0.45		
N	592		33		1	

Table 7. Comparison of institution characteristics.

Table 8. placeholder table for the kitchen sink instructor characteristics regression that will include rsquared

A Appendix Tables and Figures

B Estimation Details

B.1 Estimating Individual and Instructor Variances

We estimate the variances of instructor value-added σ^2_μ and individual error σ^2_ϵ using maximum likelihood estimation, following Gilraine et al. (2020). First, we residualize the outcomes A^*_{ijct} as in Equation 3. We model the residuals A_{ijct} as

$$A_{ijct} = \mu_j + \epsilon_{ijct}$$

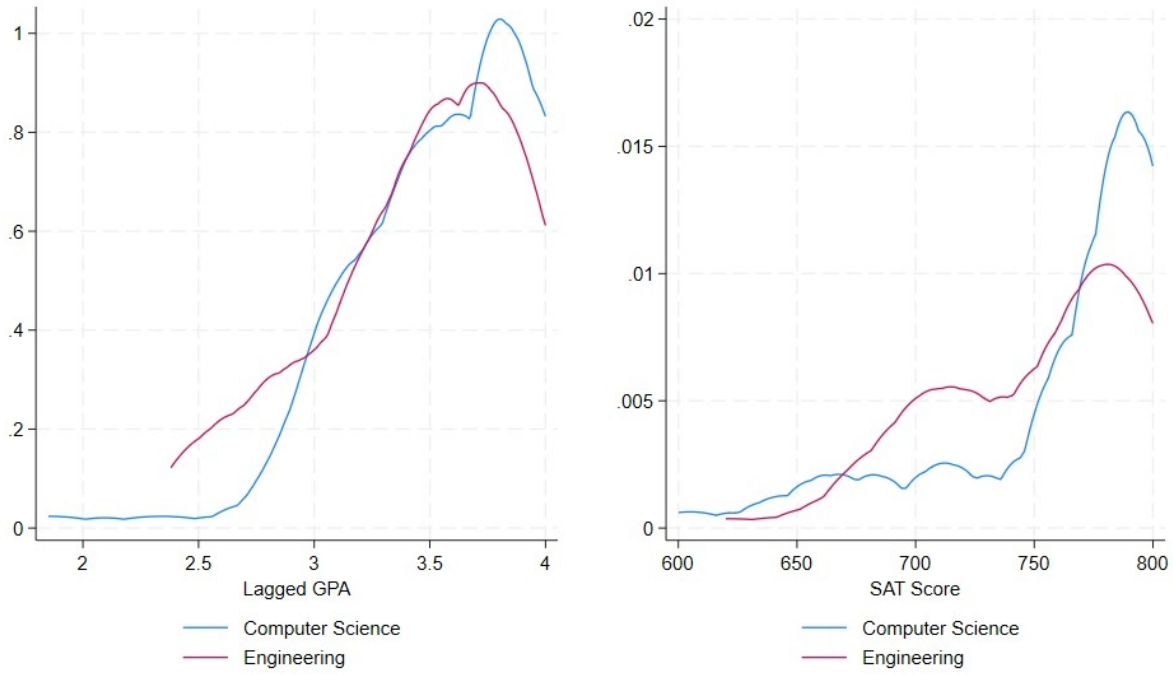
We then construct instructor-period averages $A_{jt} = \frac{1}{N_{jt}} \sum_c \sum_i A_{ijct}$, where N_{jt} is the number of students in instructor j's courses in period t. Also, denote $\mathbf{A_{jt}}$ as the vector collecting observations A_{ijct} of this set of N_{jt} students. Assuming that $\mu_j \overset{\sim}{\text{iid}} N(0, \sigma^2_\mu)$ and $\mu_j \overset{\sim}{\text{iid}} N(0, \sigma^2_\epsilon)$,

Table 9. History cluster examples

Panel C: Calculus 3	Computer Science Type (1)	Engineering Type (2)
Most Frequently Taken Courses	CS18000 CS18200 CS19100 CS19300 MA16200	ENGR16100 ENGR16200 HONR19901 HONR19902 MA16200
Last Semester GPA	3.57	3.47
SAT Score	760	751
Sorting Test p-Value	0.02	

Notes: [NTD]

Figure 3. Lagged achievement for computer science and engineering types in calculus 3 at Purdue



Notes: This figure shows smoothed kernel density plots for students in two different course history groups taking Calculus 3 in the same semester at Purdue.

Gilraine et al. (2020) show that the likelihood of the residuals takes the form

$$\mathcal{L}(A_{ijct}|\sigma_\mu^2,\sigma_\epsilon^2) = \prod_j \prod_t L_1(\mathbf{A_{jt}}|\sigma_\mu^2) L_2(A_{jt}|\sigma_\mu^2,\sigma_\epsilon^2)$$

$$L_1(\mathbf{A_{jt}}|\sigma_\mu^2) = \frac{1}{\sqrt{N_{jt}}} \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^{N_{jt}-1} \exp\left(-\sum_c \sum_i (A_{ijct} - A_{jt})^2 / 2\sigma_\epsilon^2\right)$$

$$L_2(A_{jt}|\sigma_\mu^2,\sigma_\epsilon^2) = \frac{1}{\sqrt{2\pi(\sigma_\mu^2 + \sigma_\epsilon^2/N_{jt})}} \exp\left(-\frac{A_{jt}^2}{2(\sigma_\mu^2 + \sigma_\epsilon^2/N_{jt})}\right)$$

We then maximize this likelihood numerically to obtain $\hat{\sigma}_\mu^2$ and $\hat{\sigma}_\epsilon^2$.

B.2 Empirical Bayes Estimation

Because the variance of individual error σ_ϵ^2 is much larger than the variance of value-added σ_μ^2, simple fixed effects estimates of μ_j will be affected by classical measure. Then, if we regress outcomes on value-added, such as in our forecast bias tests, coefficient estimates on value-added will be biased. To address this issue, we estimate value-added shrunk using empirical Bayes, following Kane and Staiger (2008); Chetty et al. (2014a); Bau and Das (2020), and most other modern value-added studies.

To estimate empirical Bayes estimates of value-added, we begin by residualizing the outcomes A_{ijct}^* as in Equation 3 and obtain residuals A_{ijct}. We then estimate σ_μ^2 and σ_ϵ^2 as discussed in Appendix B.1. With these estimates in hand, we construct average residuals $\overline{A}_{jt} = \frac{1}{N_{jt}} \sum_c \sum_i A_{ijct}$, where N_{jt} is the number of students taught by instructor j in period t in a given subject.

We then construct weighted sums of these residuals, using weights w_{jt}:

$$m_j = \sum_t w_{jt} \overline{A}_{jt}$$

$$w_{jt} = \frac{h_{jt}}{\sum_t h_{jt}}$$

$$h_{jt} = \frac{1}{\hat{\sigma}_\mu^2 + \frac{\hat{\sigma}_\epsilon^2}{N_{jt}}}$$

These weights up-weight the contributions of periods where instructor j teaches more students since the averages $\overline{A}_{jt}$ are more precise estimates of value-added in these periods, and down-weight contributions of periods with fewer students.

Finally, we construct empirical Bayes estimates by multiplying m_j by a shrinkage term:

$$\hat{\mu}_j = m_j \left(\frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2 + (\sum_t h_{jt})^{-1}} \right)$$

This term shrinks value-added estimates toward zero in two ways. First, it attenuates estimates more for instructors who taught fewer students total, since their estimates are less reliable than instructors who taught more students. Second, it attenuates all value-added estimates more in subjects where $\hat{\sigma}_\mu^2$ is smaller relative to $\hat{\sigma}_\epsilon^2$, since in these subjects, measurement error is more of a problem.

B.3 Jackknife Empirical Bayes Estimates

When regressing an outcome on estimates of value-added, such as in the forecast bias test in Equation 6, there may be spurious correlations if data from periods found on the left hand side of the equation are used to estimate value-added on the right hand side of the equation. To address this issue, we construct jackknife empirical Bayes value-added estimates for these types of forecast bias tests Chetty et al. (2014a).

To implement jackknife empirical Bayes value-added, we first residualize outcomes A_{ijct}^* and estimate variances as normal. Then, in the second step, we construct the weighted sums m_j and weights w_{jt} and h_{jt} leaving out period t, or both period t and $t-2$, if the forecast bias test regresses changes in outcomes on changes in value-added. This results in time-varying weighted sums m_{jt}, which we multiply by the updated shrinkage term, using the new weights h_{jt} that leave out t. Finally, we obtain jackknife empirical Bayes estimates $\hat{\mu}_{jt}$.

Though these estimates vary across time, our assumption that μ_j is fixed across time remains unchanged. The time-varying nature of $\hat{\mu}_{jt}$ is merely a statistical artifact of the estimation procedure.

B.4 Hierarchical Clustering

[NTD-clean up the way we use H_{ict} here, that's what we use for the fixed effect too]

Hierarchical clustering is an unsupervised method for grouping data. Applying this method to data requires three choices: a measure of divergence between observations, a method for measuring divergence between a groups, and a level of divergence to define the final cluster. To group students together based on their course histories, we apply hierarchical clustering to H_{ict}, which we define as the vector of indicators for each course offered at an institution, where an entry is 1 if a student has taken the course corresponding to the entry during a period $t' \leq t$ and a zero otherwise. The methods described are also used to

group students at Purdue based on their top six course preferences, similarly encoded in the vector P_{ict}, but without loss of generality, we will focus on histories.

To measure divergence between observations, we use the Jaccard index. Given a set of course histories $\{H_{ict}\}_{i=1}^{N_{ct}}$ for students in course c in period t, we calculate divergence between the histories H_{ict} and H_{jct}, with $i \neq j$:

$$g(H_{ict}, H_{jct}) = \frac{H'_{ict} H_{jct}}{H'_{ict} H_{jct} + (I - H_{ict})' H_{jct} + H'_{ict}(I - H_{jct})}$$

where I is a vector of ones. Intuitively, this measures the fraction of matches between student i and j's course histories relative to the total number of courses either student has taken. We use this measure because these vectors are very sparse.

To summarize distance between groups, we use the average linkage. This method measures the divergence between groups as the average divergence between all pairs of observations in each group. Let H^1_{ct} and H^2_{ct} be histories of groups of students in course c during period t. The average linkage between these groups is

$$G(H^1_{ct}, H^2_{ct}) = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=2}^{N_2} g(H_{ict}, H_{jct})$$

With an average linkage in hand, hierarchical clustering constructs a cluster analysis with the following algorithm:

1. Treat each observation H_{jct} as a singleton group
2. Calculate the average linkage $G(H^a_{ct}, H^b_{ct})$ between all groups $a \neq b$
3. Join the two groups with the smallest average linkage for a new set of groups
4. Repeat steps 2 and 3 until all observations are in a singleton group

This produces a large set of nested possible clusters. Finally, we choose a level of divergence to define which cluster to use. To choose a level of divergence, we calculate the mean of the levels of divergence at which each observation was first grouped and use the corresponding cluster analysis.

We chose to use the mean divergence first, to tie our hands and avoid cherry-picking, and second, to balance the trade-off between between group size and within-group similarity. Choosing a very low divergence level to form clusters results in many students being left in singleton groups, with grouped students being very similar. Those students in singleton groups are not used for estimation in the main specification. On the other hand, choosing a

high divergence level puts many students into one large group. This means that students in the large group may actually be quite dissimilar, which does not solve the sorting problem.

Table X shows summary statistics on our chosen clusters. XXX fill this in [NTD-make this table]

C Robustness Tests

C.1 Course-level [NTD] Forecast Bias Test

C.2 Including Ungrouped Students

C.3 Change in Enrollment

We test this assumption with a robustness check that regresses changes in student enrollment on changes in value-added, and find that changes in value-added within a course do not predict changes in student achievement. The full results of this test are in the appendix.

The forecast bias test described in Equation 6 relies on the assumption that $\mathbb{E}[\xi_{slt}|\Delta M_{slt}] = 0$, or that changes in student unobservables are independent of changes in value-added within a subject-level. While innocuous in K-12, this assumption could be more concerning in post-secondary education where students are able to choose when to take a course from an instructor. For example, if students could perfectly predict teaching rosters and waited to take an intermediate-level economics course from a particular instructor, this test would be biased. Therefore, we rely on the assumption that most changes in teaching rosters are unexpected by students, or that students do not react to these changes.

We examine evidence for this assumption holding by investigating how well changes in value-added predict changes in enrollment within a subject-level. Let ΔM_{slt} be the change in average value-added within a subject and level from period $t-2$ to period t, and ΔN_{slt} be the change in average enrollment for the same subject-level and period. Our robustness test regresses changes in average enrollment on changes in average value-added:

$$\Delta N_{slt} = \delta \Delta M_{slt} + \xi'_{slt}$$

An estimate of $\delta = 0$ would indicate that students are not systematically enrolling in subjects and levels where value-added is higher or lower. If students do wait to enroll in courses to have instructors that have higher (or lower) value-added, this test would have an estimate of $\delta > 0$ ($\delta < 0$).

We present results that show that students do not seem to increase or decrease their enrollment in response to changes in value-added. Figure [NTD] presents these results for

Texas. Here is a summary measure of the results [XXX NTD]