# Instructor Value-Added in Higher Education*

Merrill Warnick†
Jacob Light‡
Anthony Yim§

October 25, 2024
**[Click here for latest version]**

**Abstract**

Estimating post-secondary instructors' value-added is challenging because college students select their courses and instructors. In the absence of sound measures of value-added, universities use subjective student evaluations to make personnel decisions. In this paper, we develop a method to estimate instructor value-added at any university. The method groups together students who have previously taken similar courses and estimates value-added based on differences in outcomes for students in the same group and same course who have different instructors. Using a unique policy at a large public university in Indiana, we show that our non-experimental method controls for selection just as well as methods that exploit conditional random assignment of students to courses. We next show that our method reduces forecast bias in a wider variety of institutions using data from nearly all public universities in Texas. 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 17%. Strikingly, value-added is only weakly correlated with student evaluations. An instructor retention policy based on value-added would result in 2.7% higher earnings for students attending Texas universities.

*We are grateful for generous feedback from Jeff Denning, Caroline Hoxby, Guido Imbens, Isaac Sorkin, Jann Spiess, and participants in Stanford's Labor and Public Economics seminars. Warnick and Light acknowledge generous support from from George P. Shultz Dissertation Support Fund at SIEPR, the Institute for Research in the Social Sciences at Stanford University, and the Leonard W. Ely and Shirley R. Ely Graduate Student Fellowship through a grant to the Stanford Institute for Economic Policy Research. The results detailed in this paper do not reflect the opinions of Purdue University or the Texas Education Research Commission. All errors are our own.

†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

Teaching is central to the mission of many universities. Universities demonstrate their commitment to instructional quality by considering teaching when making important personnel decisions such as hiring, tenure, promotion, and retention. While institutions strive to provide high-quality instruction, they often lack objective measures of instructor quality. Absent objective measures, universities rely instead on subjective student evaluations. However, research indicates that these evaluations can distort teaching incentives by encouraging grade inflation (Nelson and Lynch, 1984; Eiszler, 2002) and may also reflect students' biases (Chisadza et al., 2019).

Given the limitations of subjective evaluations, quantitative measures of instructor quality, such as value-added models extensively used in K-12 education, present an appealing alternative. However, estimating value-added in higher education is complicated by substantial identification challenges. Except when applied in an experimental or randomization-based setting, value-added estimation requires that selection be on observables — i.e. the factors influencing student selection across instructors can be observed and 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 account for selection (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 selection. 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. When students' unobservable intentions are correlated with both their choice of instructor and future outcomes, like subsequent course selection or career path, conventional value-added estimates that do not address this form of selection will be biased.

This paper proposes a general method for estimating instructor value-added at many universities by augmenting the value-added model with students' "course histories." Motivated by the intuition that some students who pursue the same major have different unobserved "types" that steer them to different instructors, our approach aims to overcome bias from student selection by identifying value-added from the differences in outcomes of students who have previously taken similar classes but, for some current class of interest, have different instructors. Importantly, although the course history data our method relies on are rich, they are come from transcripts, which are necessarily maintained by every post-secondary institution. Estimates of value-added that control for course histories show that instructors impact their students' future grades and earnings.

To see the intuition of augmenting value-added with course histories, consider two stu-

dents taking Organic Chemistry with different types: one may take it as a pre-requisite for medical school, while another may take it on the path towards becoming a chemist. Informed by their intentions, students who fit these archetypes likely select different instructors if they perceive that the instructors may differentially help them obtain what they want to get out of the course. We propose that students reveal their types through the courses they have taken previously. Thus, we might distinguish "medical school types," who have previously taken Human Anatomy and Biology, from "chemist types," who have previously taken Calculus. Our method creates groups of students with similar unobservable type using hierarchical clustering based 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 otherwise unobservable differences that might bias conventional value-added estimates.[1]

In the spirit of LaLonde (1986), we begin by comparing value-added with course histories to an experimental benchmark, using data from Purdue University. At Purdue, we leverage an unusual policy that assigned students to courses randomly, conditional on submitted preferences.[2] 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 use methods that could be applied at any university. We find that course history and conditionally random value-added generate very similar rankings of instructors, with a correlation between the two methods of within-subject instructor rankings of 0.83. We also estimate value-added using lagged achievement to control for selection to document that course histories or a policy that randomizes students into classes are indeed necessary for identification. Similar to value-added in at K-12 setting, the lagged achievement value-added estimates control for observable measures of lagged student achievement, such as past GPAs and entrance exam scores, and fixed student and classmate characteristics.

Using value-added estimates from Purdue, we find that controlling for course histories reduces forecast bias substantially, relative to the standard lagged achievement model. Forecast bias often occurs when unobservably higher-ability students select higher-quality teachers. In such a case, value-added does not predict student outcomes one-for-one out of sample, and a forecast bias test will yield coefficient estimates far from 1. We assess forecast bias using two tests. The first, which we call the teaching roster changes test, is identified by year-to-year changes in teaching rosters that are unexpected by students, following Chetty

[1] 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 selection across instructors, once we restrict comparisons to be between students with similar course histories, is more plausibly random.

[2] Most prior value-added estimates for instructors in higher education use similar policies that randomize student enrollment.

et al. (2014a). The second, which we call the conditional random assignment policy test, uses value-added estimated with data from before Purdue's policy to predict student outcomes during the later period, when the policy was in place. When these bias tests are applied to lagged achievement value-added estimates — which include only the conventional controls used in K-12 value-added estimation— the estimates do not predict student outcomes out of sample, with bias test coefficients between 0.25 and 0.32. Therefore, these controls are insufficient to control for student selection. In contrast, the bias tests indicate that our value-added estimates with course histories predict student achievement out-of-sample very well, with bias test coefficients between 0.7 and 0.9. We also find that value-added with course history controls for selection approximately as well as estimates that use the conditional random assignment policy.

We use our course histories value-added method to estimate value-added to earnings and grades by applying our method to 33 public four-year universities in Texas. An appealing feature of the Texas data is our ability to link transcript data to students' earnings. This allows us to estimate value-added to earnings in addition to value-added to next-semester grades. With differing strengths and weaknesses, these two outcomes are natural starting points for value-added estimation in higher education. Semester GPA is a frequently observed measure of student ability that is of interest to both students and universities, as grades impact students' graduation rates, eligibility for certain majors, and graduate school prospects. However, GPA is internal to the university and is influenced by grade inflation and differing grading patterns across subjects. At the other end of the spectrum, earnings have a clearer connection to student welfare, are determined by the market, and are comparable across subjects and universities, but is not immediately observable for students. Many other intermediate outcomes could fit between grades and earnings.

In Texas, we find that course histories control for student selection at universities without randomization-based enrollment policies. We apply the teaching roster changes forecast bias test to estimates of value-added to grades and earnings using course histories to control for selection. For both value-added to grades and earnings, the value-added estimates predict student achievement very well: 75% of forecast bias coefficients for value-added with with course history controls are between 0.75 and 1.25, indicating that course histories address selection in this more general setting. Value-added estimates that use the lagged achievement model fail to control for selection, with the majority of forecast bias coefficients falling below 0.6.

Our estimates of value-added using course histories document that individual instructors can affect both students' grades and students' earnings. An instructor with a 1 standard deviation higher value-added increases their students' next-semester GPA by 0.13 grade

points and earnings six years post-college entry by 17%. Additionally, we find heterogeneity in the variance of value-added by subject and institution.

Similar to the K-12 literature, we find that instructor characteristics not strong predictors of value-added. Specifically, the $R^2$ in a regression of value-added on a rich set of observables is less than 0.02 for both measures. Nevertheless, a few characteristics are statistically significant predictors of value-added. For example, associate professors and contingent instructors have significantly higher value-added to earnings than full professors and assistant professors.

Finally, to assess whether value-added measures can be leveraged by universities to improve student outcomes, we estimate the earnings gains possible from a counterfactual instructor retention policy that uses value-added rather than student evaluations. To conduct this exercise, we scraped all instructor evaluations from one of the schools in the Texas sample: Texas Tech. Descriptively, we find that student evaluation scores are mildly positively correlated with value-added to GPA but are uncorrelated with value-added to earnings. Interestingly, a stronger correlate of student evaluations is not value-added, but instructor grading leniency. The correlation between evaluation scores and the average grades an instructor assigns is nearly twice the magnitude of the correlation with value-added to GPA. Given the weak or non-correlation between evaluations and value-added, there are likely potential gains to students' grades or earnings from a policy that leverages value-added in instructor retention.

We find that a counterfactual policy using value-added rather than student evaluations to make retention decisions for contingent instructors could increase student earnings. We motivate our counterfactual policy by documenting that, at Texas Tech, the likelihood that a contingent instructor is retained is roughly linear for most student evaluations, but drops discontinuously for instructors who receive evaluations in the bottom vigintile of evaluations. This discontinuity implies that the institution does indeed use student evaluations in retention decisions for contingent instructors. We then simulate a counterfactual policy that uses value-added in the way evaluations are currently used for retention and estimate earnings gains under the counterfactual policy to be in the range of 2.7% of quarterly earnings for the average student.

This project makes three primary contributions. First, we develop and validate a novel approach for estimating instructor value-added in higher education, overcoming selection issues that previously limited research to a few institutions with unique enrollment policies. The spirit of this approach may be relevant in other value-added settings where agents select on some unobservable characteristic that may be correlated with some other observable behavior. Second, we provide new evidence on the importance of instructors, demonstrating

significant variation in value-added impacts on both grades and earnings. Finally, we extend our method to a broad sample of institutions and subjects, providing insights into characteristics associated with high value-added instructors in a more nationally representative context.

This project contributes to several strands of literature. A small but growing body of research estimates instructor value-added in higher education (Hoffmann and Oreopoulos, 2009; Carrell and West, 2010; Figlio et al., 2015; Brodaty and Gurgand, 2016; DeVlieger et al., 2018). 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) leverage 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 develop a method for estimating value-added that can control for selection at the vast majority of institutions that do not use randomization or other restrictive policies to assign students to courses. An additional contribution is that, 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.[3] Our paper extends these methods to the higher education context and

[3]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).

finds that university instructors similarly affect student outcomes. Our work 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 teacher value-added to test scores on long-term outcomes like college attendance and earnings. We similarly estimate value-added to post-college earnings, but we differ by estimating the impact of instructors on earnings directly.

The rest of the paper proceeds as follows. Section 2 describes the foundational framework for value-added estimation. Section 3 describes the two panel data sources used for this project. Section 4 describes the limitations of conventional value-added models in the general higher education setting. Section 5 describes and validates our new "course histories" approach. Section 6 uses our value-added estimates to describe the characteristics of high-quality instructors. Section 7 assesses the potential for welfare improvements from instructor retention policies guided by value-added, rather than student evaluations. Section 8 concludes.

## 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 $A$.[4] 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

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

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

[4]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' next-semester GPA as a comparable academic achievement measure. To measure student outcomes more concretely, we also use student log earnings six years post-college entry.

where $X_{it}$ captures student $i$'s background characteristics, $C_{jsct}$ are characteristics of other students taking class $c$ with instructor $j$,[5] $\rho_c$ is a course fixed effect, $\lambda_t$ is a period fixed effect and $\nu_{ijsct}$ is a composite error term, which contains individual error $\epsilon_{ijct}$ and $\mu_{js}$, which is instructor $j$'s value-added to $A_{ijsct}$. 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 $\mu_j$, instructor $j$'s value-added, and the variance of the distribution of value-added, $\sigma^2_\mu$, which describes the impact of moving to a higher-quality instructor. In particular, the standard deviation of the value-added distribution $\sigma_\mu$ 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 $c$.[6] Next, we assume that both $\mu_j$ and $\epsilon_{ijct}$ are distributed normally, allowing us to use maximum likelihood estimation (MLE) to estimate variances for both distributions, following Gilraine et al. (2020).[7]

To estimate $\mu_j$, 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 $\mu_j$. 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).

Using the $A_{ijct}$, we construct three different measures of value-added, which we use for different tests because of their statistical properties. Our main measure of value-added used in all bias tests is constructed using empirical Bayes methods. Applying empirical Bayes is common in the value-added literature, since value-added estimates are subject to classical measurement error from typically large variances in the student error term, $\epsilon_{ijct}$. 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

[5] 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.

[6] 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.

[7] See Appendix C.1 for more details on this method.

distribution.[8]

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.[9]

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,[10] 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 for value-added to test scores (Kane and Staiger, 2008; Chetty et al., 2014a) and value-added to other non-test outcomes (Jackson, 2018; Petek and Pope, 2023), but there is little evidence on when this assumption holds in the higher education setting. We discuss 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. Purdue is classified as an R1 university due to its high level of research activity. Our dataset covers student transcripts from 2011 to 2023, providing

[8] For additional detail, see Appendix C.2.

[9] Appendix C.3 contains more information about jackknife estimates.

[10] Bias in these regressions arises when a regressor is correlated with the unobserved error in the model. If un-shrunk value-added estimates are used as regressors, there may be a correlation between the measurement error in these estimates and the unobserved error in the model. However, when value-added is the outcome variable, the measurement error is no longer associated with a regressor, so correlation with the unobserved error is not a concern.

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 assigned to courses through this algorithm each semester, including the Spring semesters.[11] 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.[12]

This random assignment policy allows us to construct benchmark estimates of value-added using methods which use conditional randomization to students to courses. Using randomization of students to courses in this way is the common method for estimating value-added in higher education, so we can compare course history value-added to the accepted method. Additionally, the policy 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. These transcripts are linked to the state's unemployment insurance system, allowing us to track quarterly earnings for students who remain employed in Texas.[13] 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 not done

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

[12] For more details, see Appendix B.

[13] 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.

in K-12, and explore the relationship between instructor characteristics and value-added.[14] Second, the breadth of institutions in the sample is considerable: we analyze data from 33 public universities operating continuously from 2011 to 2021.[15] 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 (NCES, 1995-2022). This diversity 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 A-1 compares other characteristics of the Texas sample and Purdue to the typical US public university.

An additional margin along which the Texas sample reflects characteristics of the average university is in its use of contingent instructors. According to faculty counts from the National Center for Education Statistics, 27% of instructors at the average Texas university are in contingent positions (e.g., instructor, lecturer), slightly more than the corresponding national value (20%). We will revisit the importance of contingent faculty in Section 7, where we consider counterfactual faculty retention policies targeted primarily at instructors on flexible contracts, such as contingent faculty.

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 Texas Tech. These evaluations, dating back as early as 2006, include overall student impressions of the instructor and 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 selection in value-added estimation for higher education. We make this point at

[14] Chetty et al. (2014b) estimate the effect of teacher value-added on income rather than estimating teacher value-added to income itself.

[15] 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.

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 selection 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 selection patterns of classmates (Altonji and Mansfield, 2018).[16] These, along with a few other background characteristics, control for selection 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 selection will be biased. For example, consider two types of students taking organic chemistry: "medical school" types take organic chemistry as a prerequisite 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.[17] 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-

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

[17] Students could be choosing instructors in other ways as well. For example, all students might choose instructors with easier grading standards. If all students selected instructors 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.

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 selection. 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 selection, 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 bias introduced when using only traditional controls versus a more robust set of controls that account for selection 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.[18] 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 selection by directly controlling for these lists of ranked preferences.

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

[18]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.

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

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.[19]

### 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 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 selection, instructor effects tend to be overstated. This happens because high-ability students are more likely to choose higher-quality instructors, making these instructors appear more effective than they truly are. Conversely, lower-quality instructors may seem worse because they are chosen by lower-ability students. As a result, the variance of the estimated value-added distribution is distorted.[20]

Following Chetty et al. (2014a), we define the forecast bias $B$ of value-added estimator $\widehat{\mu_j}$ as $1 - \alpha$, where $\alpha$ 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, $\alpha$ 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 $\psi_{ijct}$ 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

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

[20] 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 selection.

value-added distribution, but $(1 - B)\sigma_{\hat{\mu}}$.

We apply two tests for forecast bias that control for selection in different ways. The first controls for selection by using semester-to-semester changes in teaching rosters, thus differencing out unobservable differences in student selection. The second leverages Purdue's conditional random assignment policy, whereby we evaluate whether out-of-sample value-added estimates from before the policy predict student outcomes during the conditional random assignment period.

The first forecast bias test, which we call the teaching roster changes forecast bias test, or "roster test" for short, follows Chetty 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' true impacts by using real-world shifts in instructor assignments as a natural experiment.

The first forecast bias test, which we call the teaching roster changes forecast bias test, or "roster test" for short, follows Chetty 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 loads. We include include subject-by-semester fixed effects so that the identifying variation comes from changes within subject that occur across course levels within the same semester. For example, an instructor may switch from teaching the fall freshman-level course in macroeconomics to the fall senior-level course in macroeconomics.

The roster test is valid so long as changes in student unobservables are unrelated to the changes in teaching rosters. A potential violation of this assumption would occur if students time their enrollment in order to study with high value-added instructors.[21]

The roster test shows whether the method being implemented systematically over- or under-estimates instructors' value-added. In the roster test, we compare changes in the

[21] 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 D.3.

average value-added of instructors within a subject and course level to corresponding changes in average student outcomes. If the value-added estimates are unbiased, then the change that occurs when we replace a low value-added instructor with a high value-added one (or vice versa) should predict the change in average student outcomes one-for-one. However, if the value-added estimates are biased, they will not predict student outcomes one-for-one.

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:

$$\Delta A_{slt} = \overline{A_{slt}} - \overline{A_{sl,t-2}}.$$

Define $\Delta M_{slt}$ 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.

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 selection.

In contrast, value-added estimates derived from the conditional random assignment approach effectively control for selection. 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 selection 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 selection. Let $\widehat{\mu_j}^{pre}$ be value-added estimated on data from 2011-2017, before the conditional random

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

[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 1.** Teaching roster changes forecast bias test for Purdue

| | Lagged Achievement | | Conditionally Random | Course Histories | |
|---|---|---|---|---|---|
| Period: | 2018-23 | 2011-23 | 2018-23 | 2018-23 | 2011-23 |
| | (1) | (2) | (3) | (4) | (5) |
| $\Delta$ 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 |
| Conditionally Random | | | 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 selection.

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 assignment group similarity 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 predict the performance of students who move to a new school, arguing that movers do not have the information necessary to select certain instructors. In our setting, we directly control for student selection 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 selection.

The results in this section confirm that conventional controls from the K-12 value-added literature are inadequate for addressing the student selection that is prevalent in higher education. Without properly accounting for this selection, value-added estimates using these

**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 selection.

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 selection.

# 5 Augmenting Value-added with Course Histories

Unlike our previous application in Purdue, most universities do not have policies that restrict student selection across instructors. We propose a method to control for student selection 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 selection into instructors' classes poses a challenge for value-added estimation when the model fails to account for the unobservable factors driving this selection. This is because those unobservables — such as preferences, ambitions, intentions, and latent abilities — are likely correlated with students' future outcomes. Returning to our example of the two types of students who take Organic Chemistry — students who intend to go to medical school and students who intend to become chemists — if these types select different instructors, value-added estimates that do not account for this selection will be biased.

In this section, we argue that students may reveal their types through the courses they have previously taken, and that we can use these past courses to make the relevant unobservables as good as observable for the purpose of estimating value-added. We call this set of courses previously taken a student's **course history**.[24] A student's course history may in fact be a sufficient statistic for the effects of unobservable variables on selection into instructors' classes. This is a somewhat subtle point. Of course, we cannot make unobservables

[24]In the estimation, 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.

observable. What we can do, however, is focus on students within the same course history "type" so that we compare only students who have the same unobservables.

The logic of this approach is familiar within education economics. In an influential methodological paper, Dale and Krueger (2002) proposed a comparison of the earnings of students who were admitted to the same set of colleges. In their case, the unobservables were student aptitudes and preferences for various college attributes and colleges' observations of the students' qualities that are only revealed in essays, campus visits, and interviews. Two students with the same admission "portfolios" necessarily applied to the same colleges, thereby revealing their unobserved preferences, motivations, and aptitude. Furthermore, two students with the same portfolio must have been viewed as similar by admissions officers who observe essays and interviews which are unobservable to the econometrician. Thus, argued Dale and Krueger, application portfolios could plausibly make the unobservables that are relevant for college selection as good as observable. Put another way, two students with the same portfolio might be so alike on unobservables (as well as observables) that their actual college choices were plausibly random.

Similar to admissions portfolios, course histories contain rich information that may make the unobservables relevant for instructor selection as good as observable. When students choose courses, they take into account many of their own unobservables: career goals, motivation, research interests, and social networks. For example, medical school type students choose to enroll in MCAT preparation courses in Psychology and Human Anatomy that chemist type students would likely not take. Students also show that they anticipate succeeding in a course when they enroll, analogous to students applying to a college. Furthermore, students must actually succeed in a course for it to show up in their course history, analogous to colleges' admitting students. Whether or not course histories accomplish the task of embodying unobservables is an empirical question we address in Sections 5.3 and 5.4.

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.[25] We encode course histories as indicator vectors of all possible courses.[26] 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 D.2 shows that pooling all singleton students within each course and period into a single reference group has little effect on the results.

### 5.2 Common sense checks of course history groups

Hierarchical clustering based on course histories often yields student groups that align with common sense understandings of how past coursework signals future intentions. We provide three demonstrative examples of groups identified by our hierarchical clustering method. Corresponding to each example, the panels of Appendix Table A-4 list the some of the courses that appear most commonly in the course histories of these students, as well as their average GPA and SAT scores. Note that all students in these examples were in the course during the same semester.[27]

The first example is of two groups identified among students taking Organic Chemistry at Texas A&M. Two distinct groups are apparent in this course: "chemist types" and "medical school types." Students in the first group took courses typical of students on track to complete majors in Chemistry or Biomedical Engineering: Engineering Mathematics, Computational Engineering, Fundamentals of Chemistry, and courses from the Biomedical Engineering program. In contrast, students in the second group commonly took courses in the Health core, indicating these students were likely taking Organic Chemistry to meet a medical school requirement, and courses in Psychology and Sociology, courses strongly recommended for MCAT preparation. The commonly-taken courses for this second group are typical of students preparing for medical school admissions.

The second example is of groups identified among students taking Intermediate Microeconomics, again at Texas A&M. Students in the first group had previously taken courses in the business core: Marketing, Accounting, and Management Information Systems. Students in the second group were identified because they had previously taken courses in the Agricultural Economics program.[28] The clustering, therefore, distinguishes "business" and "agricultural economics" type students enrolled in Intermediate Microeconomics.

[25] For additional details about our hierarchical clustering approach, see Appendix C.4.

[26] 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.

[27] We do not report the semesters for privacy.

[28] Texas A&M has a strong program in agriculture, which was the central focus of the university when it was established as the Agricultural & Mechanical College of Texas in 1876.

The third example highlights groups of students who took Calculus 3 at Purdue. Students in the first group had most frequently taken Computer Science core courses, and are likely "computer science" type students. Students in the second group 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.

These examples reveal a form of unobservable heterogeneity in students enrolling in the same course: course histories allow us to categorize students into "types" that may reflect latent differences in motivations, ability, or work ethic. Such differences can bias value-added estimates when they correlate with how students choose instructors. To assess this, we conducted Pearson's $\chi^2$ independence tests to see if students from different course history groups systematically selected different instructors.[29] For each course, we reject the null hypothesis that course history groups and instructor choice are independent at the 10% confidence level, and for Organic Chemistry and Calculus 3, we reject the null at the 5% level.

This differential selection by course history group could not be fully controlled for by lagged achievement alone. For instance, the "business" and "agricultural economics" type students in Intermediate Microeconomics had nearly identical GPAs from the previous semester, as did the "computer science" and "engineering" students in Calculus 3. Despite small differences in entrance exam scores, the distributions of previous-semester GPA and SAT scores show substantial overlap between groups for both courses.[30] Thus, controlling only for lagged achievement without accounting for course history groups would involve comparisons across these distinct groups, likely leading to biased value-added estimates for instructors, as students in different course history groups appear to have different future intentions.[31]

While course histories address the selection challenge in these three examples, the question of whether they address selection more broadly remains yet unanswered. In the next section, we conduct forecast bias tests to show that course histories can indeed control for student selection.

### 5.3 Forecast bias estimates for course history value-added at Purdue

To test whether controlling for course histories effectively addresses bias from students' instructor choices, we apply both forecast bias tests described in Section 4.2 to value-added

[29] Appendix Table A-4 summarizes the p-values from these tests for each example.

[30] See Appendix Figures A-1 and A-2.

[31] This issue may partly stem from the limitations of GPA and SAT scores as measures of student ability, particularly in higher education. However, these are the most widely available metrics, and few institutions provide better alternatives.

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 jackknife value-added, which excludes data from period $t$ and $t-2$ the value-added estimate, 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 selection 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. Column 2 shows the results for course history value-added. With a point estimate of 0.72, course histories address unobservable student selection 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, the different measures estimate value-added for slightly different groups of students.[32] 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.

We find that the course history and conditionally random value-added methods 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-

[32]A robustness check in Appendix D.2 shows that pooling all singleton students within each course and period into a single reference group has little effect on the results.

added make very similar distinctions between low- and high-value-added instructors, course history value-added is indeed addressing student selection at Purdue.

### 5.4 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 selection 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 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.[33]

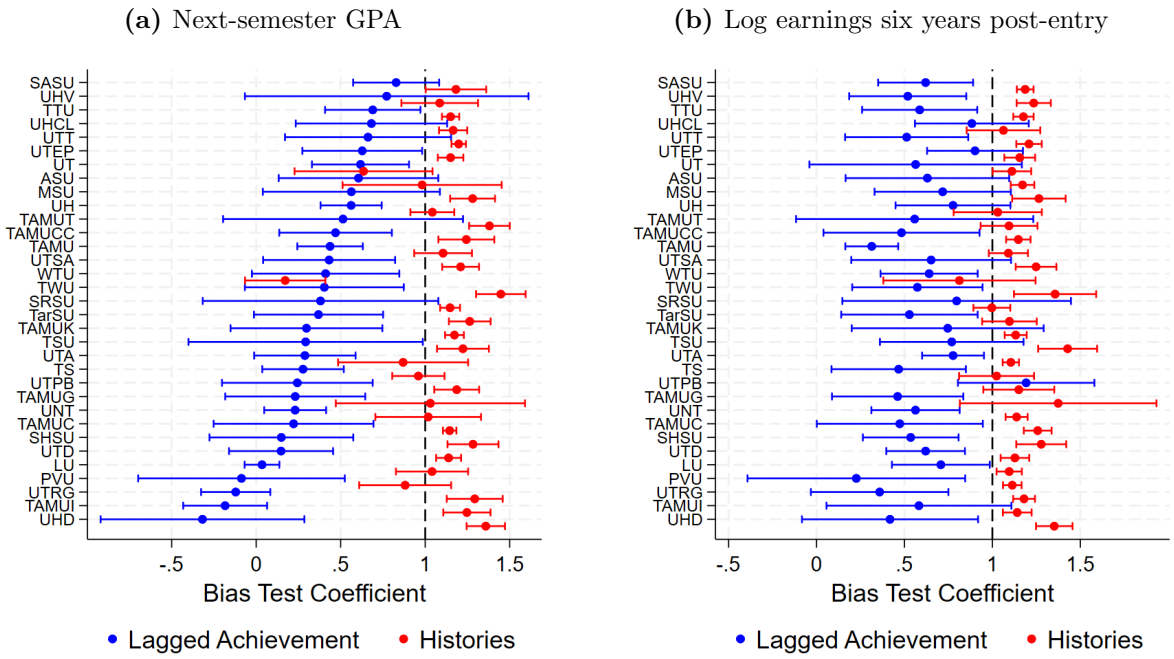
We first verify that these value-added measures are forecast unbiased. To do so, we apply the teaching roster change forecast bias test, using quasi-experimental variation in teaching rosters to identify the forecast bias coefficients. Figure 1 plots the forecast bias coefficient estimate for each university, 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 selection. 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 selection 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.

[33]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 selection.

**Table 3.** Variances of value-added distributions

| | Next-semester GPA | Log Earnings |
|---|---|---|
| | (1) | (2) |
| All | 0.018 | 0.028 |
| R1 | 0.022 | 0.036 |
| Non-R1 | 0.016 | 0.025 |

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 2010 Carnegie classification: R1 universities have "very high research activity."

# 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 that instructor characteristics explain very little of the variation in value-added. Finally, we compare our value-added estimates to students' subjective evaluations of instructor quality.

## 6.1 Instructors impact students' future outcomes

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 4.5% 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 17% six years after they enter college.[34] 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.

We find that the variance of instructor value-added varies across institutions. 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

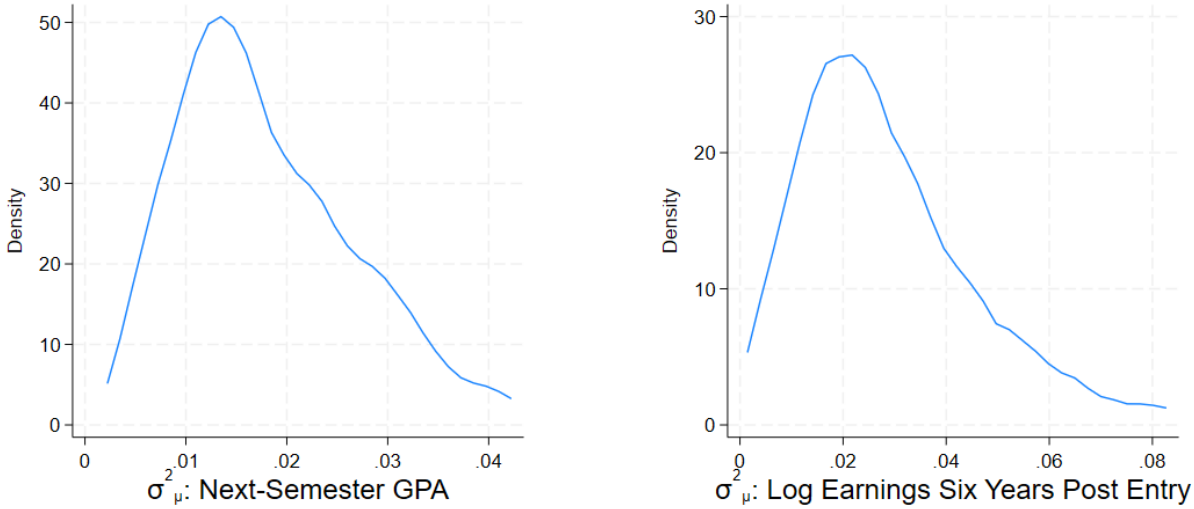
[34]This estimate includes only students who have non-zero earnings in at least one quarter during the sixth year post-matriculation. We observe earnings for more than 80% of the students in our sample. Students without earnings data likely fall into two categories: those with no wage earnings during 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.

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 both grades and earnings at R1 universities could be due to the fact that the the distributions of both next-semester GPA and log earnings six years post-entry is are more skewed at R1 institutions than non-R1 institutions. The average within institution and subject skew for next-semester GPA is $-1.47$ at R1 institutions and $-1.12$ and non-R1 institutions. The average skew for log earnings six years post-entry is $-1.43$ at R1 institutions and $-1.43$ at non-R1 institutions. The patterns for earnings is intuitive: since more 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.

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

**(a)** Next-semester GPA

**(b)** Log earnings six years post-entry




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).

Additionally, 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 0.006 and 0.038, corresponding to grade increases of 0.08 to 0.19 for moving to a 1 s.d. better instructor. Average value-added to earnings is more spread, with variances between 0.008 and 0.08, corresponding to income increases of 9% to 28% for moving to a 1 s.d. better instructor.

**Table 4.** Variances of value-added for selected subjects

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

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.[35] 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. 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). Intuitively, 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

[35]CIP stands for Classification of Instructional Programs. These identifiers from the National Center of Education Statistics allow us to compare majors across institutions even when institutions have different major prefixes.

the same cohort. The average standard deviation of log earnings for students in Education courses is 0.64, much lower than for students in Engineering or Computer Science courses (0.75 for both subjects).[36] These results 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

Fixed instructor characteristics explains little of the variation in value-added. Table A-3 shows regressions of standardized estimates of value-added on a set of instructor characteristics in Texas: rank, race/ethnicity, gender, age and salary.[37] 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. The $R^2$ in both regressions is small, with values of 0.016 for value-added to GPA and 0.009 for value-added to earnings. This result aligns with some of the value-added literature in K-12, which finds that observable characteristics do little to predict teacher value-added.[38]

Although instructor characteristics do not account for the variation in value-added, we still observe some heterogeneity in average value-added across different instructor characteristics. Table 5 summarizes the estimates from separate regressions of standardized value-added on various instructor characteristics. 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). Value-added estimates are standardized, so the interpretation of the coefficient is standard deviation difference in average value-added for a given outcome for instructors having a given characteristic relative to the base category. Additionally, for rank and race regressions, in Panels A and B, 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.

In Panel A, we highlight the regressions by academic rank. While there are no signif-

[36]Standard deviations were averages taken across institutions within a subject, weighted by number of student-course-period observations.

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

[38]Seminal work by Hanushek (1971) and Ehrenberg and Brewer (1994) suggest that characteristics of K-12 instructors are only weak predictors of student achievement, while more recent work suggests that other factors, such as principal evaluations and certifications, predict achievement (Jacob and Lefgren, 2004; Clotfelter et al., 2007). In higher education, some work has studied differences in instructor impacts between tenure-track and contingent faculty. There is little consensus in this area: previous work has found adjunct instructors to improve (Bettinger and Long, 2010; Figlio et al., 2015), reduce (Ehrenberg and Zhang, 2005), or have no effect on student achievement relative to tenure-track instructors.

**Table 5.** Heterogeneity in value-added across instructors

| | Value-added | |
|---|---|---|
| | GPA | Earnings |
| | (1) | (2) |
| A: Academic Rank | | |
| Full Professor | -0.026 | 0.019 |
| | (0.017) | (0.020) |
| Associate Professor | -0.021 | 0.071*** |
| | (0.018) | (0.021) |
| Non-Tenure Track | -0.005 | 0.047** |
| | (0.014) | (0.015) |
| N | 69,475 | 62,790 |
| F-test | 0.218 | 0.052 |
| B: Race | | |
| Asian | -0.040** | -0.030** |
| | (0.015) | (0.013) |
| Black | -0.051*** | -0.020 |
| | (0.013) | (0.015) |
| Hispanic | 0.016 | 0.008 |
| | (0.016) | (0.014) |
| N | 69,537 | 62,851 |
| F-test | 0.001 | 0.090 |
| C: Female | -0.002 | 0.013 |
| | (0.014) | (0.010) |
| N | 69,537 | 62,851 |
| D: International | -0.061*** | -0.007 |
| | (0.012) | (0.011) |
| N | 69,537 | 62,851 |

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 institution level. Value-added values are standardized, so the interpretation of the coefficient is standard deviation difference in average value-added for a given outcome for instructors having a given characteristic relative to the base category.

icant differences in value-added to GPA across ranks, we do find significant differences in value-added to earnings. On average, full professors have similar value-added to assistant professors, whereas associate professors and contingent instructors exhibit significantly higher value-added than assistant professors.[39] Additionally, an F-test rejects the null hypothesis that these coefficients are equal.

### 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 Texas Tech covering 2006-2023. 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 slightly across years and incorporate a wide variety of questions. To create a standardized measure, we focused on three types of questions that appeared in all years. Questions in the "Instructor Score" and "Course Score" categories capture students' general impressions of the instructor and the course. Questions in the "Teaching Score" 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.[40]

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 to be on a common scale. The first column shows that value-added to GPA and evaluation scores are significantly correlated. The largest coefficient indicates that on average, instructors with 1 s.d. better Teaching Score evaluations have 0.102 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

[39]These findings are consistent with Figlio et al. (2015), who report higher levels of student learning in first-semester courses taught by contingent faculty compared to tenure-track faculty at Northwestern University.

[40]For the exact questions and categorizations, see Appendix E. Additionally, we show correlations with two other categories of questions that were only available for part of our panel.

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

| | Value-added | | |
|---|---|---|---|
| | GPA | Earnings | Leniency |
| | (1) | (2) | (3) |
| A: Instructor Score | 0.082*** | -0.004 | 0.163** |
| | (0.019) | (0.022) | (0.051) |
| N | 3,204 | 2,956 | 3,531 |
| $R^2$ | 0.008 | 0.000 | 0.028 |
| B: Teaching Score | 0.102*** | -0.007 | 0.157*** |
| | (0.023) | (0.023) | (0.046) |
| N | 3,204 | 2,956 | 3,531 |
| $R^2$ | 0.012 | 0.000 | 0.024 |
| C: Course Score | 0.088*** | -0.015 | 0.169*** |
| | (0.020) | (0.025) | (0.049) |
| N | 3,204 | 2,956 | 3,530 |
| $R^2$ | 0.009 | 0.000 | 0.030 |

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 Texas Tech between 2006-2023. The specific questions related to instructor, teaching, and course quality are detailed in Appendix E. 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 subject level.

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 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 Score" category, and also score higher in overall evaluations: an instructor with 1 s.d. higher course quality scores has 0.169 s.d. higher leniency scores. Furthermore, the correlation between leniency and evaluations is stronger than the correlation between value-added to GPA and

evaluations. For each category of evaluation questions, the correlation between leniency and evaluation scores is approximately twice the magnitude of the correlation with value-added and evaluation scores. This provides suggestive evidence that students favor instructors who assign higher grades.[41]

# 7 Policy Implications

Since student evaluations are relatively uncorrelated with measures of instructor value-added, we consider policies that would make personnel decisions based on value-added rather than the current policy that relies only on student evaluations. Using student evaluations from Texas Tech, we find that evaluation scores are predictive of retention, especially for instructors in the bottom 5% of the evaluation distribution. We then benchmark the possible gains using two deselection exercises: one that replaces the bottom 5% of instructors, motivated by work in Hanushek (2009) and Chetty et al. (2014b), and one that uses the estimated relationship between evaluations and retention. Both find gains to selecting instructors on value-added to earnings relative to selecting students on student evaluations.

## 7.1 Student evaluations and retention of contingent instructors

To understand how institutions actually use student evaluations in personnel decisions, we investigate the relationship between the student evaluations an instructor receives and their probability of teaching at the same institution the next year. For this exercise, we focus on contingent instructors at Texas Tech, as contingent instructors are less likely to have longer-term employment contracts restricting termination.

We define "retention" $R_{j,s,y+1}$ of a given contingent instructor $j$ in subject $s$ as an indicator for whether that instructor is teaching a course again in subject $s$ at Texas Tech during the next academic year. We then calculate the average "Instructor Score" $S_{j,s,y}$ from student evaluations for instructor $j$ in subject $s$ during year $y$. We predict retention with the evaluation score both linearly and non-linearly, using

$$R_{j,s,y+1} = S_{j,s,y} + \lambda_s + \epsilon_{j,s,y} \tag{9}$$

$$R_{j,s,y+1} = \sum_{v=1}^{19} S^v_{j,s,y} + \lambda'_s + \epsilon'_{j,s,y}, \tag{10}$$

where $S^v_{j,s,y}$ is an indicator for instructor $j$ having an evaluation score in the $v$th vigintile

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

within subject $s$ during year $y$, with the top vigintile as the reference group. We also estimate these equations, adding controls for estimates of value-added and vigintiles of estimated value-added, respectively.

We find that student evaluation scores predict retention. Appendix Table A-5 shows results from the linear regressions from Equation 9, with and without value-added estimates. Both regressions show that evaluations significantly predict retention. An instructor with 1 s.d. better evaluation scores is 4 percentage points more likely to be retained, on average, which is about 5% of the mean, or roughly 10% of a standard deviation. Neither measure of value-added significantly predicts retention. The estimates of coefficients on value-added in Column 2 are an order of magnitude lower than the corresponding estimates for evaluation scores.

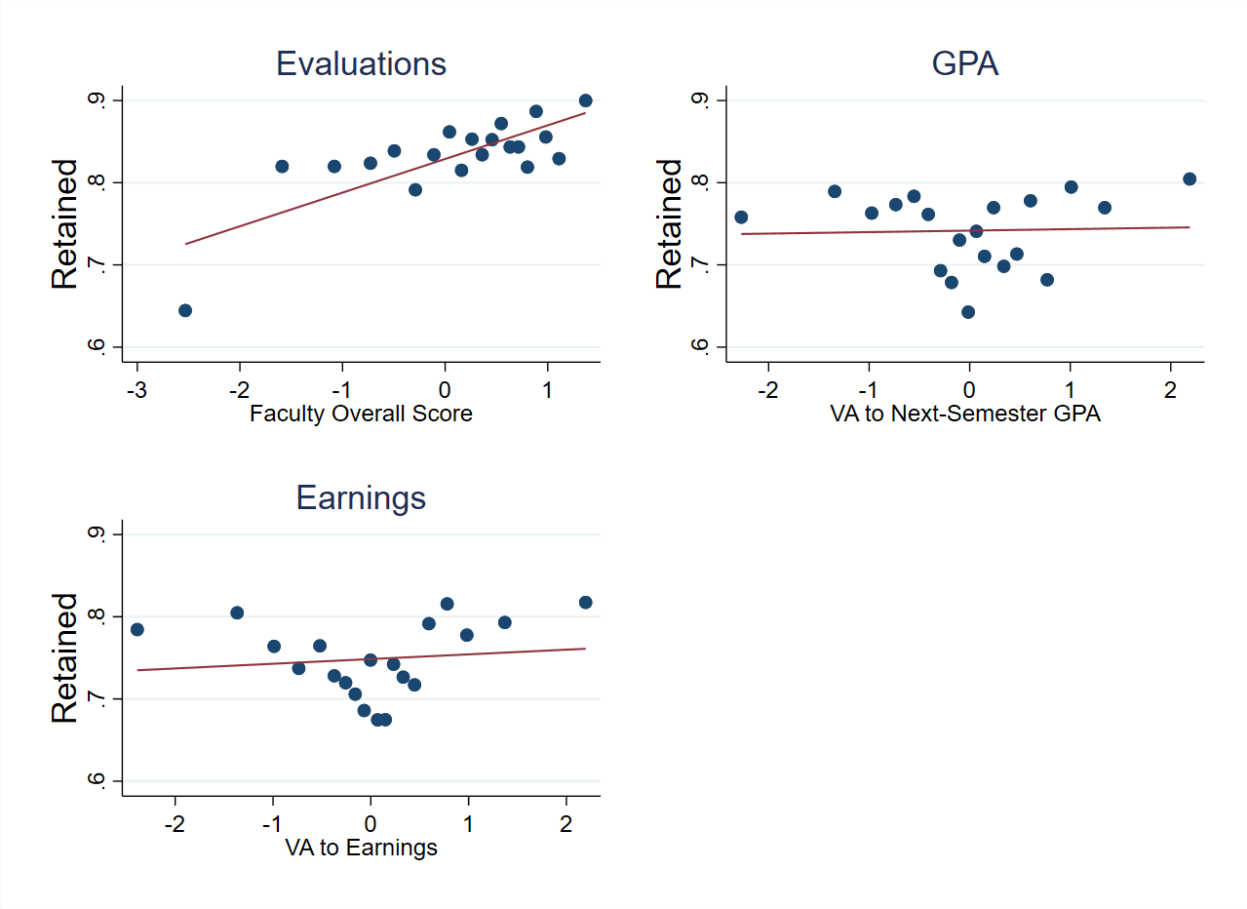
We also find that student evaluation scores and retention have a non-linear relationship. Appendix Table A-6 shows results from the vigintile regression in Equation 10. The bottom 5% of instructors is significantly less likely to be retained than other groups. Relative to the omitted top 5%, instructors in the bottom 5% are nearly 30% less likely to be retained. The bottom 5% is also less likely to be retained than the next lowest vigintile. The top left panel of Figure 3 shows a binned scatter plot of retention against evaluation scores. Above the bottom 5%, evaluations and retention have a somewhat linear relationship. However, the likelihood of retention plunges for the instructors who receive the worst evaluations. Additionally, the other two panels of Figure 3 show that value-added does not have a strong relationship with value-added.

The drop in retention rates for contingent instructors in the bottom 5% of the student evaluation distribution is notable. Moreover, deselection in this range aligns with work by Hanushek (2009) and Chetty et al. (2014b), which discuss potential gains in K-12 student achievement under policies that deselect instructors of the bottom 5% based on value-added measures. In Section 7.2, we perform a back-of-the-envelope counterfactual exercise, using the value-added distribution as a basis for deselecting instructors, rather than the current approach, to assess the potential benefits of incorporating value-added in instructor evaluations. In Section 7.3, we conduct an additional exercise that uses the estimated relationship between retention and student evaluations to assess the benefits to using value-added in evaluations of contingent instructors in a more realistic way.

### 7.2 Full deselection using value-added

To calculate the possible gains from using value-added to evaluate instructors instead of student evaluations, we conduct back-of-the-envelope calculations in the spirit of following Chetty et al. (2014b). This exercise is motivated both by the deselection suggestion in

**Figure 3.** Retention rates and instructor quality at Texas Tech




Notes: Binned scatter plots of retention indicators against three measures of instructor quality for contingent instructors at Texas Tech: evaluations, value-added to GPA and value-added to earnings. Each point plots the average retention against average instructor quality measure within vigintiles of instructor quality measures. Additionally, we show the best fit lines from a regression of retention on the instructor quality measure. All variables are residualized on subject and year.

Hanushek (2009) and by our empirical finding that, at Texas Tech, instructors in the bottom 5% of the evaluation distribution do indeed seem to be retained at much lower rates.

First, we calculate the average per-student increase in earnings six years post-college entry from replacing an instructor in the bottom 5% of the value-added to earnings distribution with a mean instructor. The ingredients of this calculation are as follows: assuming a normal distribution of value-added, an instructor in the bottom 5% is, on average, 2.063 standard deviations from the mean instructor; the standard deviation of the value-added to log earnings distribution is 0.17; and the median of quarterly earnings in our sample is

$8,638. We use median earnings instead of mean earnings to limit the effects of outliers. The average per-student increase in earnings six years post-college entry from replacing an instructor in the bottom 5% is:

$$G = 2.063 \times \$8,638 \times 0.17 = \mathbf{\$}3,029$$

or, about 35% of median student earnings. These gains are substantial.[42]

In order to investigate how de-selecting certain groups of instructors could impact earnings, we conduct an additional back-of-the-envelope calculation that uses subject-institution-specific variances of value-added. First, we identify each instructor in the bottom 5% of the value-added distribution within their subject and institution. We then calculate the gains in earnings from replacing six groups of instructors with the mean instructor, within each subject and institution: replacing all instructors in the bottom 5%, and replacing instructors of each rank separately who are in the bottom 5% of instructors, for contingent instructors, assistant professors, associate professors, full professors, and all Tenure Track instructors.

Table 7 reports results from this exercise. Both income and gains in income from replacement are highly skewed distributions, so Panel A, which presents gains from replacing all bottom 5% instructors, shows the median gain. The median increase in quarterly earnings six years post-entry from replacing all bottom vigintile instructors is $\$2,644$, or about 30% of median earnings. This substantial gain is primarily driven by two factors. First, the estimate is driven by our relatively large estimate of the variance of the distribution of value-added to earnings. Second, most students take at least one course from a bottom 5% instructor, so the proposed policy impacts many students.

Though most students have at least one bottom 5% instructor, when we split up the policy to replace instructors of different types, the median earnings gain is 0 in every case. To limit the effects of outliers for Panel B, we report the winsorized mean of affected students, scaled by the fraction of affected students, to measure the average gains from replacing the relevant group of bottom 5% instructors. We find that the largest gains are from replacing contingent instructors. The average gain from replacing all bottom 5% contingent instructors is $3,850$, which is about 42% of the mean.[43] Replacing instructors in the bottom 5% of assistant, associate, and full professors would result in smaller gains.

[42]Note that in this section, we treat estimated value-added as true value-added. Accounting for estimation error encompasses scaling estimates by their reliability, and would shrink the estimates of gains somewhat. In the next section, we address this issue in another exercise.

[43]Note that if we calculate the gains from replacing all instructors in this way rather than report the medians, the gains are larger than the gains from replacing only tenure-track instructors as shown in Row 1 of Panel B.

**Table 7.** Average gains for deselection of the bottom 5%

| | Average Gain (1) | Percent Gain over Average Earnings (2) |
|---|---|---|
| **A: Medians** | | |
| Replace All | \$2,644 | 30.61% |
| **B: Means** | | |
| Replace All | \$6,182 | 68.43% |
| Replace Contingents | \$3,850 | 42.62% |
| Replace Assistants | \$1,155 | 12.79% |
| Replace Associates | \$1,478 | 16.36% |
| Replace Full Professors | \$1,511 | 16.72% |
| Replace Tenure Track | \$2,749 | 30.43% |

Notes: Estimates are average gains to replacing instructors in the bottom 5% of the value-added distribution within each of the given categories. Column 1 shows average dollar gains from the replacements. Column 2 shows the percent gain of the row over average earnings, which has a median of \$8,638 and mean of \$9,033. Means in Panel B were constructed by averaging winsorizing gains for students with non-zero gains, then multiplying by the fraction of students with non-zero gains.

### 7.3 Partial de-selection using value-added

A central insight from the analysis in Section 7.1 is that the university's retention policies are constrained, as many instructors cannot be dismissed due to tenure or other long-term contracts. Even among contingent instructors with low student evaluations, the likelihood of retention is still greater than 50%. Given the frictions in the academic labor market, it may not be feasible — or even desirable — to replace 5% of instructors annually. Within these constraints, this section explores the potential gains from an alternative retention program that maintains the current number of annual replacements but bases retention decisions on instructors' value-added rather than student evaluations. By design, such a policy would lead to modest improvements in students' earnings.

The exercise in this section focuses on Texas Tech. We begin by predicting the probability that a contingent instructor $j$ will be retained in year $y$. We construct this probability in two ways: first, by using the prediction given from our estimates of Equation 10 to show a baseline for if the institution actually used the predicted evaluation policy, and second, by applying the estimated model to vigintiles of value-added to earnings instead of to vigintiles of evaluations. This gives us two probabilities of retention: $\hat{R}_{j,y}(SQ)$, the "status quo" probability of retention and $\hat{R}_{j,y}(VA)$, the "value-added" probability of retention. To account for differing retention rates, the regressions and predictions included year and subject fixed

effects. For instructors that cannot be replaced by the policy (tenure-track instructors, teaching assistants, and first-time instructors) we assign both retention probabilities as 1.

The model gives a probability that an instructor is retained at each value of their student evaluation score. We estimate average earnings changes using the predicted retention probabilities for the full range of student evaluation scores, as well as a more plausible policy that only uses these retention probabilities for replacing the bottom 5% of instructors.

Given these retention probabilities, the value-added standardized within a subject, $\mu^z_{j,s}$, the subject-specific variance of value-added to earnings, $\sigma^2_{\mu,s}$, and student $i$'s earnings six years post-entry, $Y_i$, the predicted gains $G_i(a)$ for student $i$ under replacement regime $a$ are:[44]

$$G_i(a) = \sum_{t=1}^{T_i} \sum_{j=1}^{N_i} \left( (1 - \hat{R}_{j,y}(a)) \times -\mu^z_{j,s} \times \sigma_{\mu,s} \times Y_i \right).$$

Note that these gains are still infeasible, in the sense that we treat $\hat{\mu}^z_{j,s}$ as true value-added for instructor $j$ instead of estimated value-added. We also estimate more feasible gains to the policy by scaling each instructor's value-added estimate by the reliability of their estimate, $r(j,s)$, from the empirical Bayes estimation.[45] This results in a gain of

$$FG_i(a) = \sum_{t=1}^{T_i} \sum_{j=1}^{N_i} \left( r(j,s) \times (1 - \hat{R}_{j,y}(a)) \times -\mu^z_{j,s} \times \sigma_{\mu,s} \times Y_i \right).$$

Table 8 shows average infeasible and feasible gains for these policies. Columns 1 and 2 show changes from the "status quo" policy. In fact, under a policy that replaces all instructors based on the evaluation retention probabilities (i.e. allowing the probability of retention to vary across the full distribution of evaluation scores, regardless of whether their score is in the bottom 5% of scores overall), students' average earnings would be *lower* than actual student earnings, on average. However, because the status quo is an approximation of the true policy that the university uses, these changes are small in magnitude, at about 0.2% of the mean. Applying the evaluation-based retention policy but only changing retention probabilities for instructors with evaluations in the bottom 5% would increase students earnings on average, but the gains are again small in magnitude at about 0.1%.[46]

On the other hand, replacing instructors based on value-added would increase student earnings substantially. Columns 3 and 4 report gains from a policy that replaces contingent instructors with probability proportional to their relative position in the distribution of

[44] Note that we treat the gains additively to simplify the calculation.

[45] Appendix C.2 gives the exact formula

[46] We do not report feasible gains or losses from the evaluation-based retention policy, but they would be even smaller in magnitude than the feasible gains.

value-added to earnings. Panel A shows that the gains from the infeasible replacement for all contingent instructors using value-added would increase average student earnings by about $300, or by about 3% of mean earnings. Furthermore, we find that replacing only the bottom 5% of instructors would have nearly as large of an effect. Replacing the bottom 5% of instructors using the infeasible policy predicts an average earnings increase of roughly 2.6%. The feasible policy that treats value-added as an estimate predicts very similar gains as the infeasible policy. Panel B shows that the gains to replacing all contingent instructors based on the retention policy but using value-added would increase earnings by about 2.7%, and the gains to replacing only the bottom 5% are 2.2%.

One limitation of this exercise is that even our estimates of feasible gains from replacing instructors using value-added ignore the fact that these value-added estimates use data from the whole sample. A true feasible policy exercise would restrict to using data from previous periods to predict retention, as is the case for evaluations. Our earnings value-added estimates are not ideal for this exercise, as the data needed to estimate value-added, earnings six-years post entry, do not materialize until up to six years after an instructor teaches a course with a student.

There are two alternative strategies for a feasible exercise. The first is to change the university's objective: students with higher grades are more likely to graduate, so a university might be interested in using estimates of instructor value-added to next-semester GPA to evaluate instructors. A second possibility is to use a measure of value-added that is correlated with value-added to earnings. The gains to using such a measure would have smaller effects on earnings, proportional to the correlation between the two measures.

## 8 Conclusion

Instructors play an important role for the production of human capital in post-secondary education. However, estimates of post-secondary instructor impacts have, thus far, been limited to a small set of unique institutions with uncommon enrollment policies. Furthermore, these impacts have generally been measured in terms of within-course standardized assessments for a small set of courses. Due to the identification and measurement challenges associated with estimating value-added, most universities use subjective student evaluations to assess instructor quality and make personnel decisions.

In this paper, we propose and validate a non-experimental method for estimating instructor value-added that can be applied broadly in higher education. We show that students' "course histories" reveal otherwise unobservable information about student types that, when controlled for, substantially reduces forecast bias in value-added estimates. After addressing

**Table 8.** Average gains for policy-based deselection at Texas Tech

| | Evaluations | | Value-added to Earnings | |
|---|---|---|---|---|
| | Avg Gain | % Gain | Avg Gain | % Gain |
| | (1) | (2) | (3) | (4) |
| **Panel A: Infeasible** | | | | |
| Replace All Contingent | -$23 | 0.238% | $301 | 3.117% |
| Replace Bottom 5% Contingent | $12 | 0.124% | $250 | 2.586% |
| **Panel B: Feasible** | | | | |
| Replace All Contingent | | | $260 | 2.690% |
| Replace Bottom 5% Contingent | | | $209 | 2.164% |

Notes: Estimates are average gains to replacing contingent instructors at Texas Tech using coefficients from a regression of retention on student evaluation vigintiles with subject-year fixed effects. Panel A shows infeasible gains, which treat value-added estimates as true value-added and Panel B shows feasible gains, which scale by the reliability for each instructor. Columns 1 and 2 show gains for using predictions generated by the actual regression that predicts retention with evaluations, and 3 and 4 show gains from replacing vigintiles of evaluations with vigintiles of value-added to earnings in the prediction. Columns 1 and 3 show average gains in dollars while even columns show percentage gains over mean earnings at Texas Tech, which was $9,651. Row 1 of both panels show gains from replacing all contingent instructors according to the rule and Row 2 of both panels shows gains from replacing instructors only in the bottom 5% of the value-added distribution.

the identification and measurement challenges, we document large variation in instructor impacts on students within a subject. Specifically, a student who moves to a 1 s.d. better instructor within a course would improve their next-semester grades by 0.13 points and increase their income by 17. Furthermore, we document that student evaluations and instructor value-added to earnings six years post-college entry are essentially uncorrelated, highlighting a shortcoming of the use of evaluations in making personnel decisions. We estimate that a policy to deselect contingent instructors which uses value-added to earnings in place of student evaluations could increase student earnings by 2.7%.

Our method, which uses students' past and present course choices to control for selection to instructors, provides new opportunities for institutions to evaluate their instructors and opens new avenues for researchers to investigate the higher education production function. For example, in the paper, we focused our attention to instructor effects averaged across all students. Future research could investigate heterogeneous instructor effects by race, gender, or ability.[47]

We view our two outcomes, next-semester GPA and income, as spanning a wide variety of outcomes which could be useful to students and institutions. At one end, semester GPA

[47] Eastmond et al. (2023) and Bates et al. (2024) are two examples of this type of heterogeneous estimation in K-12.

provides measures of student academic ability that are related to graduation and available in real time. However, GPAs are likely a noisy measure of student ability, and reflect both students' performance in future courses and characteristics of courses students select in the future (such as grading norms in the subject a student selects). At the other end, income is an outcome tied to well-being and ability that is widely comparable, but is realized with such a long lag that it may not be desirable for making short-term personnel decisions. We expect that our method can be applied to other outcomes of interest in higher education that lie between GPA and earnings, such as major choice or time to degree.[48] These more intermediate outcomes could both provide institutions with metrics of instructor quality that both are correlated with earnings and give insight into the mechanisms behind the instructor impacts we document in this paper.

[48]For example, Jackson (2018) provides evidence in K-12 that the same strategies to estimate value-added to standardized test scores can produce unbiased estimates of value-added to non-cognitive outcomes.

## 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.

Michael Bates, Michael Dinerstein, Andrew Johnston, and Isaac Sorkin. Teacher labor market policy and the theory of the second best. Technical report, August 2024. URL https://www.nber.org/papers/w29728. NBER Working Paper No. 29728, Accepted in the Quarterly Journal of Economics.

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.

Eric P. Bettinger and Bridget Terry Long. Does Cheaper Mean Better? The Impact of Using Adjunct Instructors on Student Outcomes. *The Review of Economics and Statistics*, 92(3):598–613, August 2010. URL https://ideas.repec.org/a/tpr/restat/v92y2010i3p598-613.html.

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.

Thibault Brodaty and Marc Gurgand. Good peers or good teachers? evidence from a french university. *Economics of Education Review*, 54:62–78, 2016. ISSN 0272-7757. doi: https://doi.org/10.1016/j.econedurev.2016.06.005. URL https://www.sciencedirect.com/science/article/pii/S0272775715300091.

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.

Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6):673–682, 2007. ISSN 0272-7757. doi: https://doi.org/10.1016/j.econedurev.2007.10.002. URL https://www.sciencedirect.com/science/article/pii/S0272775707000982. Economics of Education: Major Contributions and Future Directions - The Dijon Papers.

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.

Tanner S. Eastmond, Nathan J. Mather, Michael David Ricks, and Julian Betts. Effect heterogeneity and optimal policy: Getting welfare added from teacher value added. Technical report, UCSD, 2023. Working Paper, September 8, 2023.

Ronald G. Ehrenberg and Dominic J. Brewer. Do school and teacher characteristics matter? evidence from high school and beyond. *Economics of Education Review*, 13(1):1–17, 1994. ISSN 0272-7757. doi: https://doi.org/10.1016/0272-7757(94)90019-1. URL https://www.sciencedirect.com/science/article/pii/0272775794900191.

Ronald G. Ehrenberg and Liang Zhang. Do tenured and tenure-track faculty matter? *Journal of Human Resources*, XL(3):647–659, 2005. ISSN 0022-166X. doi: 10.3368/jhr.XL.3.647. URL https://jhr.uwpress.org/content/XL/3/647.

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.

David N. Figlio, Morton O. Schapiro, and Kevin B. Soter. Are Tenure Track Professors Better Teachers? *The Review of Economics and Statistics*, 97(4):715–724, 10 2015. ISSN 0034-6535. doi: 10.1162/REST_a_00529. URL https://doi.org/10.1162/REST_a_00529.

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 Hanushek. Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 61(2):280–288, 1971. ISSN 00028282. URL http://www.jstor.org/stable/1817003.

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.

Florian Hoffmann and Philip Oreopoulos. Professor qualities and student achievement. *The Review of Economics and Statistics*, 91(1):83–92, 2009. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/25651319.

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. The impact of teacher training on student achievement. *Journal of Human Resources*, XXXIX(1):50–79, 2004. ISSN 0022-166X. doi: 10.3368/jhr.XXXIX.1.50. URL https://jhr.uwpress.org/content/XXXIX/1/50.

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.

Robert J LaLonde. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, 76(4):604–620, September 1986. URL https://ideas.repec.org/a/aea/aecrev/v76y1986i4p604-20.html.

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.

Tomáš Müller, Roman Barták, and Hana Rudová. Iterative forward search: Combining local search with maintaining arc consistency and a conflict-based statistics. In *1st International Workshop on Local Search Techniques in Constraint Satisfaction*, page 1. Citeseer, 2004.

Tomáš Müller, Keith Murray, et al. Comprehensive approach to student sectioning. *Annals of Operations Research*, 181(1):249–269, 2010.

Kevin J Mumford, Richard Patterson, and Anthony Lokting Yim. College course shutouts. Technical report, CESifo, 2024.

NCES. Integrated postsecondary education data system (ipeds), 1995-2022. Accessed: 2022-12-01.

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.

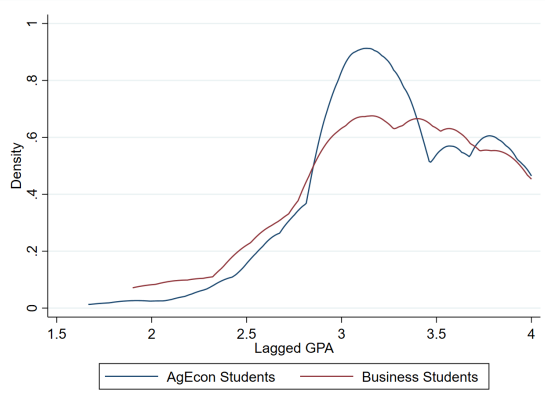
| | Population | | Texas Sample | | Purdue |
|---|---|---|---|---|---|
| | mean | sd | mean | sd | |
| Enrollment | 11,639 | 11,823 | 18,107 | 15,327 | 40,555 |
| Admit rate | 0.78 | 0.18 | 0.83 | 0.15 | 0.53 |
| SAT-equivalent: 25 pctile | 954 | 144 | 975 | 130 | 1,180 |
| SAT-equivalent: 75 pctile | 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 |
| Contingent instructors % | 0.20 | 0.14 | 0.27 | 0.11 | 0.13 |
| 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 | | 1.00 |

**Table A-1.** Comparison of institution characteristics.

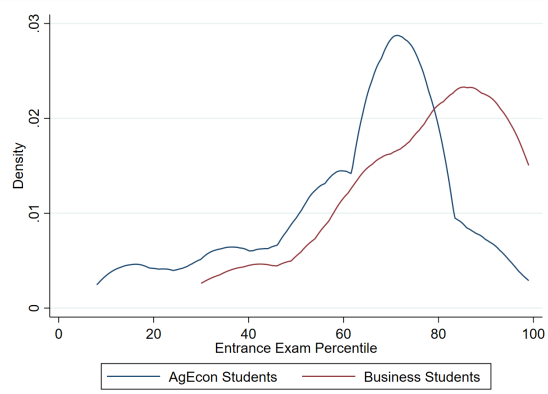
# A Appendix Tables and Figures

**Figure A-1.** Lagged achievement for business and agricultural econ types in intermediate micro a Texas A&M

**(a)** Lagged GPA




**(b)** Entrance Exam Percentile



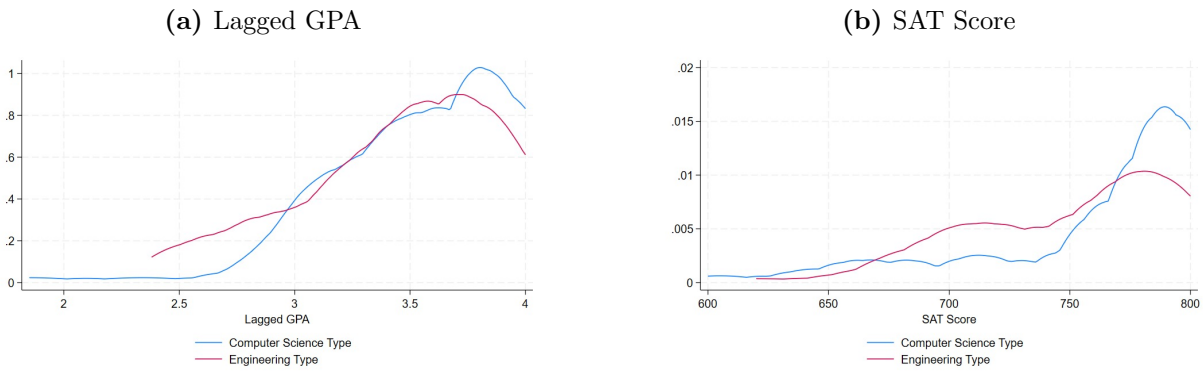

Notes: This figure shows smoothed kernel density plots for students in two different course history groups taking Intermediate Micro in the same semester at Texas A&M.

**Table A-2.** Student Characteristics

| | All | | R1 | | Non-R1 | |
|---|---|---|---|---|---|---|
| | mean | sd | mean | sd | mean | sd |
| Next-Semester GPA | 2.96 | 0.90 | 3.06 | 0.85 | 2.87 | 0.95 |
| Last Semester GPA | 2.99 | 0.72 | 3.07 | 0.69 | 2.90 | 0.75 |
| Admissions Exam Percentile | 52.97 | 16.85 | 62.01 | 16.99 | 43.60 | 16.70 |
| Log Earnings Six Years Post-Entry | 8.86 | 0.73 | 8.92 | 0.75 | 8.80 | 0.72 |
| Has Income Six Years Post-Entry | 0.72 | 0.44 | 0.69 | 0.46 | 0.75 | 0.43 |
| Bachelor's Attainment | 0.70 | 0.43 | 0.71 | 0.42 | 0.68 | 0.44 |
| Age | 21.53 | 4.33 | 21.13 | 3.71 | 21.94 | 4.96 |
| Female | 0.54 | 0.47 | 0.50 | 0.48 | 0.58 | 0.46 |
| Hispanic | 0.33 | 0.42 | 0.32 | 0.43 | 0.35 | 0.40 |
| Black | 0.13 | 0.29 | 0.09 | 0.28 | 0.17 | 0.31 |
| Asian | 0.10 | 0.26 | 0.14 | 0.31 | 0.06 | 0.20 |
| Number of Students in Section | 68.93 | 60.94 | 89.28 | 85.96 | 47.85 | 35.03 |

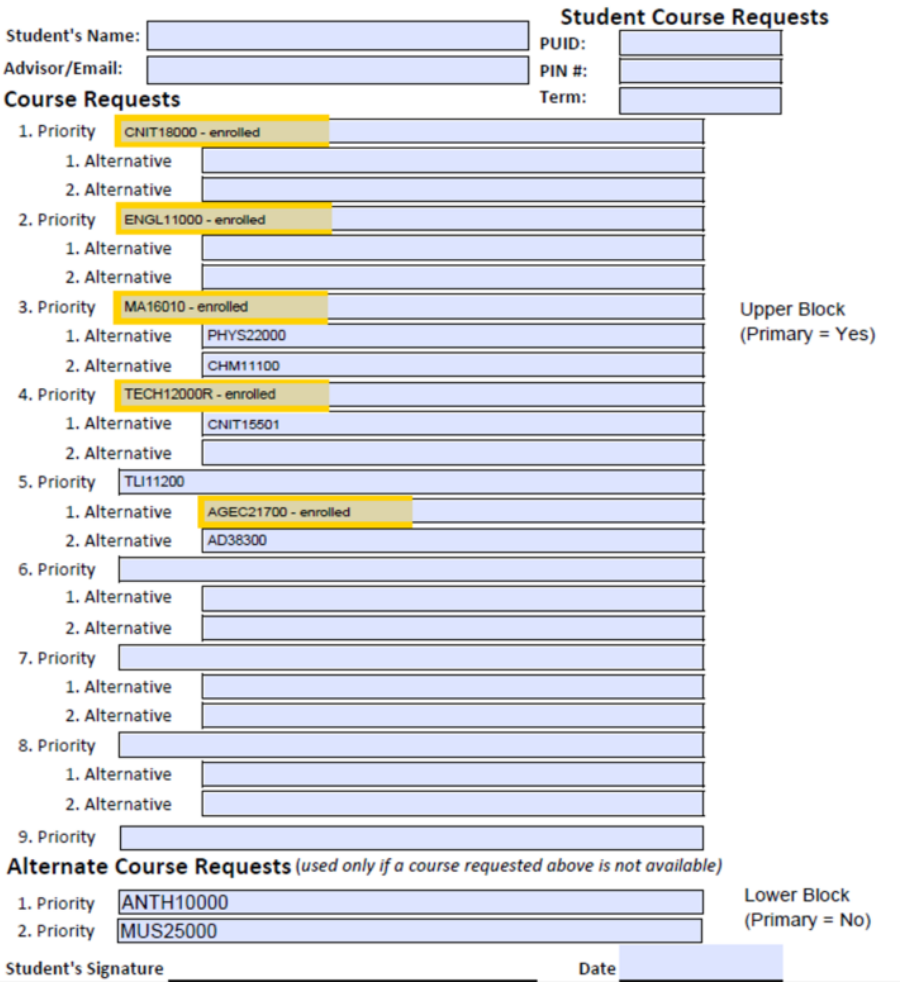
Notes: This table shows summary statistics for most of our standard control variables for students in Texas. This table takes observation-weighted averages across institutions and subjects. The first two columns use data from all institutions. The next columns split universities by their 2010 Carnegie classification: R1 universities have "very high research activity."

**Figure A-2.** 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.

**Figure A-3.** Course Enrollment Request Form at Purdue



**Student Course Requests**

Student's Name:
Advisor/Email:
PUID:
PIN #:
Term:

**Course Requests**

1. Priority: CNIT18000 - enrolled
   1. Alternative:
   2. Alternative:
2. Priority: ENGL11000 - enrolled
   1. Alternative:
   2. Alternative:
3. Priority: MA16010 - enrolled
   1. Alternative: PHYS22000
   2. Alternative: CHM11100
4. Priority: TECH12000R - enrolled
   1. Alternative: CNIT15501
   2. Alternative:
5. Priority: TLI11200
   1. Alternative: AGEC21700 - enrolled
   2. Alternative: AD38300
6. Priority:
   1. Alternative:
   2. Alternative:
7. Priority:
   1. Alternative:
   2. Alternative:
8. Priority:
   1. Alternative:
   2. Alternative:
9. Priority:

Upper Block (Primary = Yes)

**Alternate Course Requests** *(used only if a course requested above is not available)*

1. Priority: ANTH10000
2. Priority: MUS25000

Lower Block (Primary = No)

Student's Signature _____ Date

# B Purdue's Course Registration Algorithm

As discussed by Mumford et al. (2024), Purdue University assigns undergraduate courses using a distinctive course registration algorithm developed by Müller et al. (2010). This algorithm processes students' ranked course preferences as input data to generate schedules for the entire student body. It prioritizes fulfilling primary course requests while minimizing reliance on alternative requests . Each request is weighted according to the following formula:

$$weight(a \in dom(R)) = 0.9^{prior(R)} \times 0.5^{alt(a)}, \tag{11}$$

where *prior(R)* is the ranking of the requested course, and *alt(a)* represents alternate course preferences. To illustrate, consider the course enrollment request form in Figure A-3. The weight for a first-choice course with no alternatives is 0.9, while second and third choices receive progressively lower weights. The algorithm solves this allocation problem using Iterative Forward Search (Müller et al., 2004), and imposes a higher penalty for rejecting priority-only requests compared to those with alternatives.

The course assignment process follows four key constraints:

1. **Seat limits:** Each course section has a fixed number of seats, though some, such as online courses, may have no limit.

2. **Overlapping sections:** Students cannot be enrolled in overlapping course sections.

3. **Distance conflicts:** Sections scheduled too close together geographically are avoided unless there is more than a 20-minute gap between them.

4. **Course reservations:** Some courses reserve spots for students with specific majors.

# C Estimation Details

## C.1 Estimating variances of instructor value-added and individual error

We estimate the variances of instructor value-added $\sigma^2_\mu$ and individual error $\sigma^2_\epsilon$ 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 $\mu_j \tilde{\text{iid}} N(0, \sigma_\mu^2)$ and $\mu_j \tilde{\text{iid}} N(0, \sigma_\epsilon^2)$, 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 $\hat{\sigma}_\epsilon^2$.

### C.2 Empirical bayes estimation

Because the variance of individual error $\sigma_\epsilon^2$ is much larger than the variance of value-added $\sigma_\mu^2$, simple fixed effects estimates of $\mu_j$ 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 $\sigma_\mu^2$ and $\sigma_\epsilon^2$ as discussed in Appendix C.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 the reliability for instructor $j$'s estimate:

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

The reliability 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. We use this reliability factor in constructing our feasible policy gains in Section 7.3.

### C.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 $\mu_j$ is fixed across time remains unchanged. The time-varying nature of $\hat{\mu}_{jt}$ is merely a statistical artifact of the estimation procedure.

### C.4 Hierarchical clustering

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 $\mathbf{H_{ict}}$[49], 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 $\{\mathbf{H_{ict}}\}_{i=1}^{N_{ct}}$ for students in course $c$ in period $t$, we calculate divergence between the histories $\mathbf{H_{ict}}$ and $\mathbf{H_{jct}}$, with $i \neq j$:

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

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 $\mathbf{H^1_{ct}}$ and $\mathbf{H^2_{ct}}$ be histories of groups of students in course $c$ during period $t$. The average linkage between these groups is

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

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

1. Treat each observation $j$ as a singleton group
2. Calculate the average linkage $G(\mathbf{H^a_{ct}}, \mathbf{H^b_{ct}})$ 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.

[49] We use the bold font $\mathbf{H_{ict}}$ to refer to the full vector of past and current courses for student $i$ in course $c$ during period $t$, while $H_{ict}$ refers to the course history similarity group fixed effect

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 selection problem.

Table A-7 shows summary statistics on the course history similarity groups in Texas. The mean number of students per course was about 11 students, with 5 students per group on average. Most of the groups of students were groups of two. However, more than half of students were in groups of at least 15.

# D Robustness Tests

## D.1 Course-level teaching roster changes forecast bias test

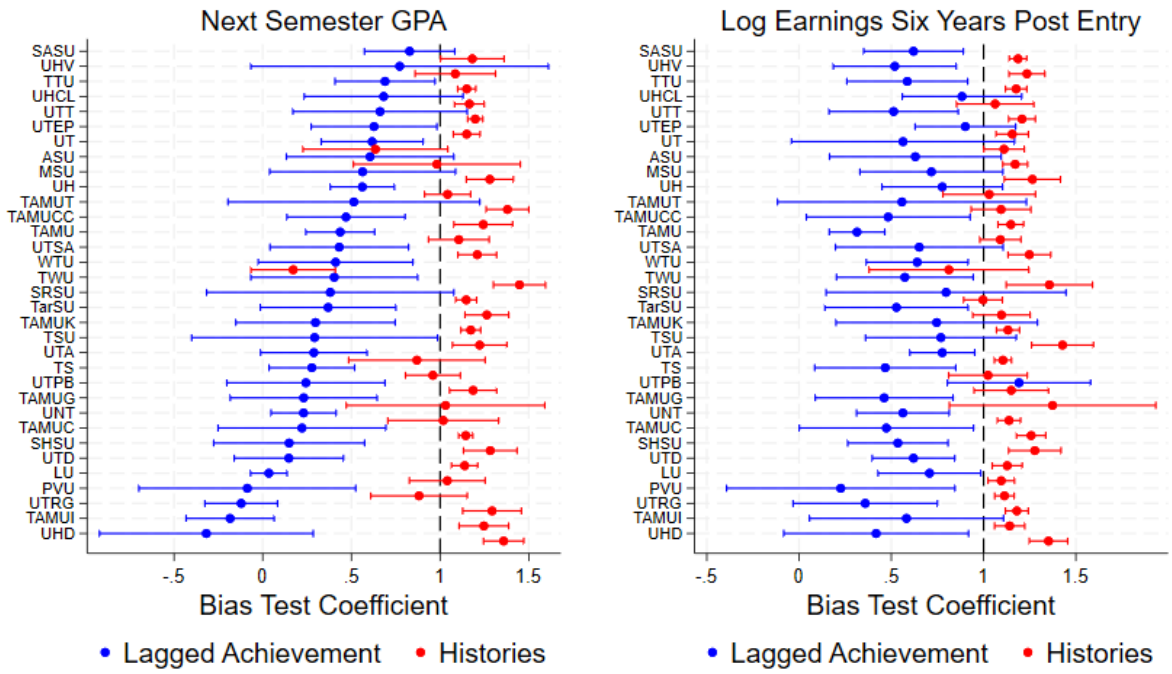
In the teaching roster changes forecast bias, we used changes in value-added to predict changes in student outcomes within a subject-course level cell. An alternative is to conduct this same test, but using course averages instead. Figure A-4 presents results from course teaching roster changes forecast bias tests for all Texas universities in our sample. Similar to the course level forecast bias tests shown previously, using course histories to control for unobservable student selection greatly reduces bias, especially relative to estimates that use only lagged achievement to control for selection. In fact, estimates with lagged achievements only may be more forecast biased than suggested by the subject-course level tests.

## D.2 Including ungrouped students

After applying hierarchical clustering to students' course histories, roughly 30% of student-course-period observations were left in singleton groups. We call these students "ungrouped". Our main results were estimated by excluding ungrouped students. Here, we include ungrouped students in the estimation procedure by putting all such students in the same course and period into a course history similarity group. Effectively, ungrouped students become the reference group in each course-period for the history group fixed effects in the residualization in Equation 3.

Figure A-5 shows forecast bias tests from all Texas institutions for value-added estimates to next-semester GPA and log earnings six years post-entry that include ungrouped students. Similar to estimates using only grouped students, estimates with ungrouped students

**Figure A-4.** Course 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-course level fixed effects. Observations are at the course-period level. Standard errors are clustered at the period-subject level. An estimate closer to 1 indicates better controls for selection.

substantially reduce forecast bias relative to estimates that use only lagged achievement to control for student selection.

Table A-8 shows average variances across Texas instituions of value-added estimates that include ungrouped students. The variance of value-added to both GPA and earnings are very close to the variances estimated using only grouped students, indicating that our sample restriction does not change the interpretation of our results meaningfully.

### D.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 $\Delta M_{slt}$ be the change in average value-added within a subject and level from period $t-2$ to period $t$, and $\Delta N_{slt}$ 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:

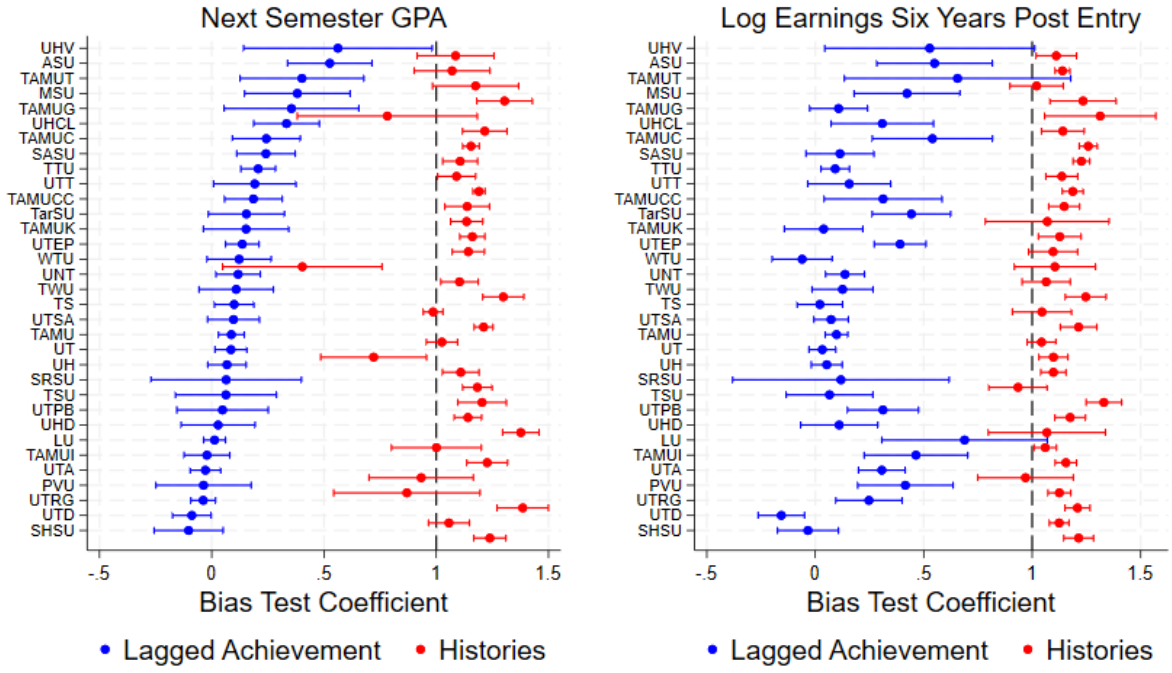
$$\Delta N_{slt} = \delta \Delta M_{slt} + \xi'_{slt}$$

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 $\delta > 0$ ($\delta < 0$).

Figure A-6 presents these results for Texas. The coefficients indicate that at nearly all universities, changes in instructor value-added do not predict changes in enrollment. Nearly all estimates have confidence intervals containing zero, and most estimates are very near zero.

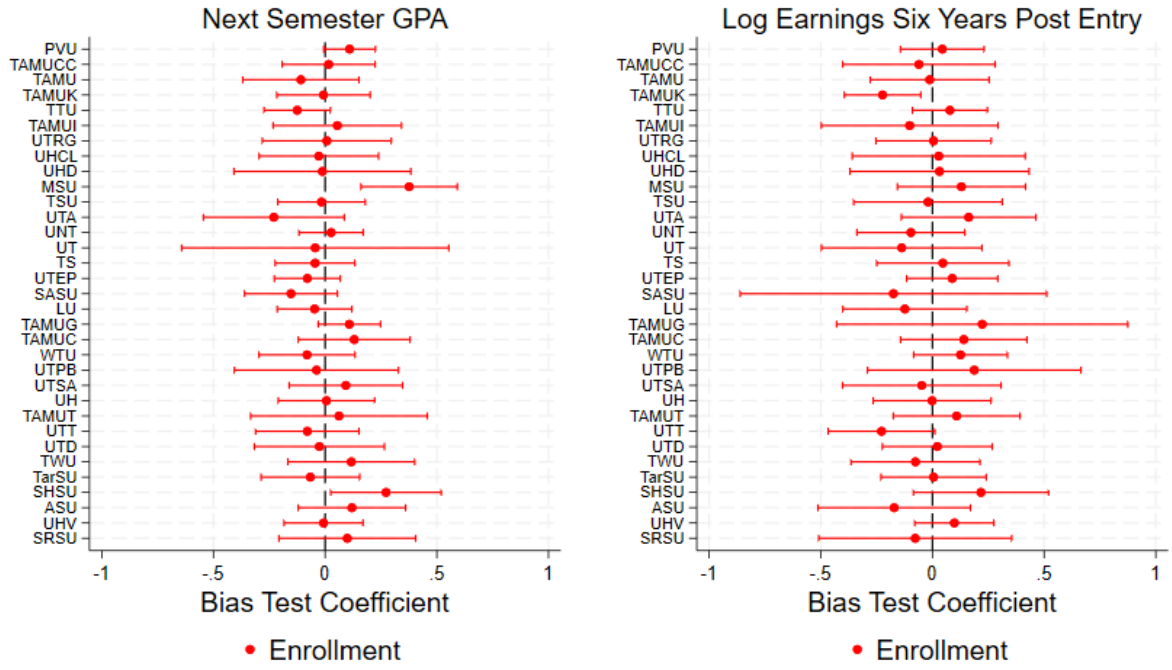
The results of this test find no evidence that students increase or decrease their enroll-

**Figure A-5.** Teaching roster changes forecast bias test for Texas universities with ungrouped students




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 selection.

**Figure A-6.** Teaching roster changes enrollment test




Notes: The teaching roster changes enrollment to assess whether changes in student enrollment are predicted by shifts in instructor value-added, with estimates regressing changes in enrollment 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 0 indicates that students do not systematically respond to changes in value-added.

ment in response to changes in value-added. This suggests that students do not time their enrollment to be with high- or low value-added instructors, suggesting that the teaching roster changes forecast bias tests are valid.

## E Student Evaluation Questions

Table A-9 shows the categorization of questions from Texas tech student evaluations. Students gave scores from 1 to 5 in response to each question, where 5 indicated that the statement was very true for the instructor. We then took averages across these responses to form evaluation scores. Note that we collected two additional categories: the "soft skills" score, that encapsulates questions concerning the instructors kindness and availability, and

the "fair course" score, which incorporates questions concerning the overall fairness of the course. We excluded these categories from the main regressions and policy evaluation because they were introduced halfway through our panel. Table A-10 shows bivariate regressions of value-added and leniency on all five evaluation scores. Again, value-added to GPA is significantly correlated with both of these evaluation scores and value-added to earnings is not. For these categories, leniency is even more highly correlated than the overall and teaching score categories.

**Table A-3.** Heterogeneity in value-added across instructors

| | VA to GPA (1) | VA to Earnings (2) |
|---|---|---|
| Full Professor | -0.550*** | 0.654** |
| | (0.116) | (0.287) |
| Associate Professor | -0.532*** | 0.688** |
| | (0.110) | (0.289) |
| Assistant Professor | -0.500*** | 0.604** |
| | (0.111) | (0.275) |
| Non-Tenure Track | -0.493*** | 0.622** |
| | (0.100) | (0.257) |
| Asian | -0.033** | -0.030** |
| | (0.015) | (0.013) |
| Black | -0.050*** | -0.023 |
| | (0.013) | (0.015) |
| Hispanic | 0.016 | 0.000 |
| | (0.016) | (0.013) |
| Female | -0.006 | 0.010 |
| | (0.015) | (0.011) |
| International | -0.055*** | -0.002 |
| | (0.013) | (0.013) |
| Log Average Salary | 0.047*** | -0.052* |
| | (0.011) | (0.026) |
| Age | 0.002 | 0.011 |
| | (0.007) | (0.006) |
| Fraction Upper Level | -0.070** | 0.028 |
| | (0.030) | (0.022) |
| N | 68,903 | 62,359 |
| $R^2$ | 0.016 | 0.009 |

Notes: Estimates are from separate regressions of value-added on instructor characteristics. Regressions control for subject and institution, with observations at the instructor level. Standard errors are clustered at the institution level. Value-added values are standardized, so the interpretation of the coefficient is standard deviation difference in average value-added for a given outcome for instructors having a given characteristic relative to the base category.

**Table A-4.** History cluster examples

| Panel A: Organic Chemistry | Chemist Type (1) | Medical School Type (2) |
|---|---|---|
| Commonly Taken Courses | BMEN101<br>BMEN253<br>CHEM119<br>ENGR102<br>MATH151<br>VTPP434 | CHEM120<br>HLTH210<br>HLTH236<br>HLTH240<br>PSYC107<br>SOCI205 |
| Last Semester GPA | 3.33 | 3.67 |
| Admissions Test Percentile | 86 | 59 |
| Selection Test p-Value | 0.04 | |
| Panel B: Intermediate Micro | Business Type | Agricultural Econ Type |
| Commonly Taken Courses | ACCT229<br>ACCT230<br>BUSN101<br>ECON202<br>ISTM210 | AGEC117<br>AGEC217<br>AGEC314<br>AGEC340<br>POLS207 |
| Last Semester GPA | 3.30 | 3.32 |
| Admissions Test Percentile | 76 | 62 |
| Selection Test p-Value | 0.06 | |
| Panel C: Calculus 3 | Computer Science Type | Engineering Type |
| Commonly Taken Courses | CS18000<br>CS18200<br>CS19100<br>CS19300<br>MA16200 | ENGR16100<br>ENGR16200<br>HONR19901<br>HONR19902<br>MA16200 |
| Last Semester GPA | 3.57 | 3.47 |
| SAT Score | 760 | 751 |
| Selection Test p-Value | 0.02 | |

Notes: This table compares popular courses and summary statistics for selected course history similarity groups in Organic Chemistry (Panel A) and Intermediate Microeconomics (Panel B) at Texas A&M and Multivariable Calculus at Purdue (Panel C). The selection test p-value comes from a Pearson's $\chi^2$ test of independence. For Panel A and B that come from Texas A&M, we show percentile of entrance exam score. Panel C shows average SAT score for the Purdue students.

**Table A-5.** Linear retention of contingent instructors

| | Retained | |
|---|---|---|
| | No Controls | VA Controls |
| | (1) | (2) |
| Instructor Score | 0.041*** | 0.042*** |
| | (0.008) | (0.009) |
| N | 4,213 | 3,902 |

Notes: Estimates are from regressions of a retention indicator of contingent instructor scores from student evaluations with subject and year fixed effects. Column 2 controls for vigintiles of value-added to next-semester GPA and earnings. Observations are at the instructor-subject-year level. Standard errors are clustered at the subject level.

**Table A-6.** Non-linear retention of contingent instructors

| | Retained | |
|---|---|---|
| | No Controls (1) | VA Controls (2) |
| Instructor Score Vigintile 1 | -0.289*** | -0.262*** |
| | (0.032) | (0.038) |
| Instructor Score Vigintile 2 | -0.175*** | -0.136** |
| | (0.036) | (0.041) |
| Instructor Score Vigintile 3 | -0.157*** | -0.121** |
| | (0.038) | (0.049) |
| Instructor Score Vigintile 4 | -0.176*** | -0.164*** |
| | (0.037) | (0.045) |
| Instructor Score Vigintile 5 | -0.170*** | -0.152*** |
| | (0.037) | (0.044) |
| Instructor Score Vigintile 6 | -0.173*** | -0.156** |
| | (0.039) | (0.049) |
| Instructor Score Vigintile 7 | -0.143*** | -0.124** |
| | (0.030) | (0.037) |
| Instructor Score Vigintile 8 | -0.209*** | -0.202*** |
| | (0.038) | (0.044) |
| Instructor Score Vigintile 9 | -0.157*** | -0.146** |
| | (0.038) | (0.044) |
| Instructor Score Vigintile 10 | -0.169*** | -0.137** |
| | (0.032) | (0.041) |
| Instructor Score Vigintile 11 | -0.144*** | -0.134*** |
| | (0.028) | (0.031) |
| Instructor Score Vigintile 12 | -0.148*** | -0.142*** |
| | (0.028) | (0.035) |
| Instructor Score Vigintile 13 | -0.158*** | -0.142*** |
| | (0.027) | (0.033) |
| Instructor Score Vigintile 14 | -0.118*** | -0.094** |
| | (0.026) | (0.033) |
| Instructor Score Vigintile 15 | -0.137*** | -0.121*** |
| | (0.026) | (0.033) |
| Instructor Score Vigintile 16 | -0.157*** | -0.144*** |
| | (0.028) | (0.036) |
| Instructor Score Vigintile 17 | -0.126*** | -0.111*** |
| | (0.025) | (0.031) |
| Instructor Score Vigintile 18 | -0.127*** | -0.107** |
| | (0.028) | (0.035) |
| Instructor Score Vigintile 19 | -0.104** | -0.097** |
| | (0.036) | (0.042) |
| N | 4,213 | 3,902 |

Notes: These regressions are similar to those in Table A-5, but use vigintiles of instructor scores from student evaluations. Estimates are relative to the top vigintile of instructor scores. Column 2 controls for vigintiles of value-added to GPA and earnings.

**Table A-7.** Course history similarity group statistics

| | Mean (1) | Median (2) |
|---|---|---|
| Groups per Course | 10.892 | 4 |
| Students per Group | 5.193 | 2 |
| Fraction Ungrouped | 0.299 | . |
| Fraction in Groups of $\geq 15$ | 0.505 | . |

Notes: This table shows statistics on the course history similarity groups for our Texas sample. The units of observation for rows 1 and 2 were course-periods. The units of observation for rows 3 and 4 were student-course-periods.

**Table A-8.** Variances of value-added distributions with ungrouped students

| Outcome Measure | $\sigma^2_\mu$ |
|---|---|
| Next Semester GPA | 0.014 |
| Log Earnings Six Years Post Entry | 0.023 |

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.

**Table A-9.** Evaluation Question Categorization

| Category | Question Text |
|---|---|
| Instructor Score | Overall this instructor was effective<br>Overall, the instructor was an effective teacher |
| Teaching Score | The course objectives were specified and followed by the instructor<br>The instructor demonstrated knowledge of the subject<br>The instructor stimulated student learning<br>The instructor presented the information clearly<br>The instructor emphasized the major points and concepts |
| Course Score | Overall this course was a valuable learning experience<br>Overall, this course was a valuable learning experience |
| Soft Skills Score | The instructor treated all students fairly<br>The instructor treated all students with respect<br>The instructor welcomed and encouraged questions and comments<br>The instructor was available for consultation during office hours or by appointment |
| Fair Course Score | Expectations were clearly stated either verbally or in the syllabus<br>The testing and evaluation procedures were fair<br>The workload was appropriate for the hours of credit |

Notes: Categorization of student evaluation questions from Texas Tech. Some wordings of questions changed slightly during our panel of evaluations.

**Table A-10.** Comparison of instructor value-added to student evaluations

| | Value-added | | |
|---|---|---|---|
| | GPA | Earnings | Leniency |
| | (1) | (2) | (3) |
| A: Soft Skills Score | 0.091** | 0.000 | 0.235*** |
| | (0.034) | (0.023) | (0.033) |
| N | 1,841 | 1,839 | 2,052 |
| B: Fair Course Score | 0.125*** | 0.003 | 0.260*** |
| | (0.031) | (0.021) | (0.037) |
| N | 1,841 | 1,839 | 2,052 |

Notes: Estimates are from separate bivariate regressions of instructor and course quality measures, based on student evaluations, on instructor value-added or leniency. Evaluations Texas Tech. The specific questions related to instructor, teaching, and course quality are detailed in Appendix E. 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 subject level.