Instructor Value-Added in Post-Secondary Education*

Merrill Warnick†
Jacob Light‡
Anthony Yim§

March 3, 2026

Abstract

Estimating post-secondary instructors' value-added is challenging because college students select their courses and instructors. In the absence of value-added measures, 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 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 method account 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 0.078 points, and earnings six-years after college entry by 0.083 log points. Descriptive evidence suggests that nearly half of instructor's effect on earnings operates through future major choice. Strikingly, value-added is only weakly correlated with student evaluations. Deselection of the bottom 5% of instructors using value-added to earnings instead of student evaluations could result in 3% higher earnings six-years after college entry for students attending Texas universities.

*We are grateful for generous feedback from Jeff Denning, Brigham Frandsen, Caroline Hoxby, Guido Imbens, Lars Lefgren, Isaac Sorkin, Jann Spiess, participants in Stanford's Labor and Public Economics seminars and Applied JMC Showcase. Warnick and Light acknowledge generous support from George P. Shultz Dissertation Support Fund at SIEPR, the Institute for Research in the Social Sciences at Stanford University, the Graduate Opportunity Fund in the School of Humanities and Sciences, 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,University of Georgia. mw68751@uga.edu
‡Hoover Institution, Stanford University. jdlight@stanford.edu
§Department of Economics, Brigham Young University. anthony_yim@byu.edu

1 Introduction

Value-added methods are widely used in economics to measure how heterogeneous agents contribute to outcomes in settings ranging from education to health care and labor markets (e.g., Abowd et al., 1999; Doyle Jr et al., 2010; Chetty et al., 2014a). In these applications, whether estimating teacher effectiveness, physician quality, or manager productivity, identification relies on a selection-on-observables assumption: conditional on rich covariates, exposure to different agents must be as good as random. Quasi-experimental variation in agent assignment is therefore central to credible estimation. In primary education, for example, classroom assignment rules and institutional features often generate plausibly quasi-random exposure to teachers, enabling researchers to identify teachers' causal impacts on student achievement and long-term outcomes.[1] This literature documents substantial heterogeneity in teacher effects, with meaningful consequences for economic mobility and human capital formation.

However, value-added methods can be severely biased when researchers cannot observe, and therefore cannot control for, the characteristics that govern individuals' selection of agents. In higher education, this challenge is pronounced: students exercise substantial discretion in choosing both courses and instructors, and these choices are shaped by intentions, abilities, career goals, and information sets that are unobserved by the econometrician. When such unobservables correlate with both instructor choice and student outcomes, standard value-added estimators confound selection with causal effects. As a result, despite the centrality of postsecondary instructors in shaping skill acquisition and labor market preparation, there is limited evidence on how much instructors actually matter for students' academic and economic trajectories. The few existing estimates come from rare institutional environments that generate random or algorithmic assignment of students to instructors (e.g., Carrell and West, 2010; Figlio et al., 2015; DeVlieger et al., 2018). These settings are informative but atypical; they do not provide a generalizable approach for estimating instructor impacts at scale.

The absence of credible, scalable measures of postsecondary instructor quality has first-order implications for universities and policymakers. Institutions routinely make high-stakes personnel decisions, such as hiring, promotion, tenure, and retention, using subjective student evaluations in lieu of objective performance measures. A large body of research shows that such evaluations can be distorted by grade inflation (Nelson and Lynch, 1984; Eiszler, 2002) and reflect systematic biases related to instructor gender, race, and other attributes

[1] In practice, assignment is rarely purely random, but controlling for lagged achievement appears sufficient in many K-12 contexts (Kane and Staiger, 2008; Chetty et al., 2014a).

(Chisadza et al., 2019). Without reliable information about instructors' contributions to economically meaningful measures of students' human capital accumulation or labor market success, universities risk rewarding the wrong behaviors and misallocating teaching talent.

This paper proposes a general method for estimating instructor value-added at many universities by augmenting the value-added model with students' "course histories." Intuitively, some students who pursue the same major have different unobserved "types" that steer them to different instructors. Our approach follows this intuition and overcomes bias from student selection in value-added estimation by comparing 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 come from transcripts, which are necessarily maintained by every post-secondary institution. Estimates of value-added that control for course histories show that instructors differ in their ability to impact their students' future grades and earnings.

To see the intuition of augmenting value-added with course histories, consider two students 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 because instructors may differentially help them obtain what they want to get out of the course. We propose that students reveal their types through their past course choices. 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 types using hierarchical clustering based on their course histories. By only using comparisons within these groups to identify value-added, we are able to control for otherwise unobservable differences that might bias conventional value-added estimates.[2]

In the spirit of LaLonde (1986), we begin by comparing value-added with course histories to a conditionally random benchmark, using data from Purdue University. At Purdue, we leverage an unusual policy that assigned students to courses randomly, conditional on submitted preferences.[3] 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

[2]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 restricting comparisons to be between students with similar course histories makes instructor selection more plausibly random.

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

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 a 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 cases, 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 three tests. The first, which we call the individual test, is a simple out-of-sample forecasting exercise, where we test how well leave-one-out estimates of value-added predict individual student outcomes. The second, which we call the prediction tests, checks how much our value-added estimates are biased due to potential confounding observables omitted in the estimation step. The last, , 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. Of the three, the final test has the strongest identification, leveraging actual conditional randomization coming from Purdue's assignment policy. 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 forecast bias estimates of 0.52 on the policy test and of 0.83 to 0.94 on the individual test. 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 estimated bias of 0.1 to 0.16 in Purdue, and average bias near 0 across universities in Texas. 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 show that our course histories value-added method is generalizable to other settings by estimating value-added at nearly all public universities in Texas. Texas is an ideal setting to test portability because of the size and variety of its university system. In Texas, we estimate value-added at 33 universities which enroll nearly 10% of all four-year university students in the US. These universities span a large range of university types, from large flagship R1 institutions to smaller regional public schools. Our data in Texas allows us to estimate

value-added to next-semester GPA and, importantly, to post-college earnings.[4] Existing work measuring earnings impacts of high-quality instructors typically measures earnings impacts mediated by student achievement. Ours is one of the first papers to measure value-added to earnings directly.

In Texas, we find that course histories control for unobservable student selection at universities without randomization-based enrollment policies. We apply the three forecast bias tests to estimates of value-added to grades and earnings. For both value-added to grades and earnings, the value-added estimates predict student achievement very well nearly all forecast bias coefficients for value-added with course history controls are very close to 1 in magnitude, 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 more than half of forecast bias coefficients falling below 0.6.

Our estimates of value-added using course histories document that individual instructors differ in their ability to affect both students' grades and students' earnings. An instructor with 1 standard deviation higher value-added increases their students' next-semester GPA by 0.078 grade points and earnings six-years post-matriculation by 0.083 log points, suggesting that instructors have substantial impacts on their student's grades and earnings.

Similar to the K-12 literature, we find that instructor characteristics are not strong predictors of value-added. In a regression of value-added on a rich set of observables, the R^2 is less than 0.02 for value-added to both outcomes. This suggests that value-added is new information to the institution not found in CVs or resumes. 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.

We find suggestive evidence that instructors affect student's grades and earnings through their future course and major choices. To test for mechanisms, we compare Shorrocks-Shapley decompositions controlling for past characteristics and instructor fixed effects to similar decompositions controlling for downstream outcomes, like course choice and major choice. We find that about half of instructor's contribution to next-semester grades is ex-

[4]Though there are many possible outcomes of interest for value-added estimation, next-semester GPA and earnings are natural starting points in higher education because they have different strengths and weaknesses. Semester GPA is a frequently observed measure of student ability that is of interest to both students and universities, since grades impact students' graduation rates, eligibility for certain majors, and graduate school prospects. However, GPA is internal to the university and is influenced by differing grading patterns across time, subject, and instructor. 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 are not immediately observable for students. Many other intermediate outcomes could fit between grades and earnings.

plained by future course and instructor choice, and about half of an instructor's contribution to earnings is explained by graduating major.

Finally, to assess whether value-added measures could 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 counterfactual policies using value-added rather than student evaluations to make retention decisions for contingent instructors could increase student earnings. We motivate our policies 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 relationship implies that the institution does indeed use student evaluations in retention decisions for contingent instructors. We simulate the results of a mass deselection of the bottom 5% policy, following Hanushek (2009), and find large gains for students. We then conduct a more realistic exercise, which uses the relationship between evaluations and retention to motivate a policy that uses value-added instead. We estimate the earnings gains under the realistic 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 value-added settings, both inside and outside of education, where agents select on some unobservable characteristic that may be correlated with some other observable behavior. Second, we extend our method to a broad sample of institutions and subjects, showing that it addresses selection in a more nationally representative context. Third, we provide new evidence on the importance of instructors, demonstrating significant variation in value-added impacts on both grades and earnings.

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; Shi, 2025). 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 which do not use randomization or other restrictive policies to assign students to courses.[5] 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.

Contemporaneous work by Shi (2025) examines the long-run impacts of high value-added instructors in developmental courses at community colleges, using the same Texas dataset as our study. While the institutional settings and identification strategies differ, both papers find that instructors matter for students' long-term outcomes and that observable instructor characteristics explain little of this variation.[6]

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

[5]We do make use of a policy at Purdue, but only to validate our generalizable method and establish a ground truth for value-added.

[6]Shi (2025) focuses on first-year students in developmental (remedial) courses at community colleges, a setting chosen in part to mitigate concerns about student-instructor matching. In contrast, we study a broader set of instructors at four-year institutions and address selection by conditioning on students' fixed characteristics across multiple courses. Shi estimates instructor value-added as residual pass rates and relates this to long-term outcomes; we estimate value-added directly with respect to those outcomes. Differences in effect sizes across the two papers may reflect institutional differences — such as greater earnings variance among four-year graduates compared to two-year graduates — and the lower baseline completion rates for two-year Texas students, leaving few students to unlock the wide-ranging returns to completion.

teachers in primary and secondary schools have causal impacts on student outcomes across a variety of settings.[7] Our paper extends these methods to the higher education context and finds that university instructors similarly affect student outcomes. 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. Our estimates of value-added to earnings are the first of their kind in both the K-12 and the higher education literature.

The rest of the paper proceeds as follows. Section 2 describes the foundational framework for value-added estimation and discusses identifying assumptions of our approach using course histories. Section 3 describes the two panel data sources used for this project. Section 4 describes estimation strategies for value-added. Section 5 validates our measures by conducting forecast bias tests. Section 6 uses our value-added estimates to document variation in value-added across higher education instructors and investigates mechanisms behind that variation. Section 7 compares student evaluations and value-added, and assesses the potential for earnings increases from instructor retention policies guided by value-added. Section 8 concludes.

2 Model and Identification

This section develops the economic framework for estimating instructor value-added in higher education. We begin by specifying the data generating process for student achievement and defining our estimands of interest. We then describe the standard identification strategy used in K-12 value-added estimation and explain why it is likely to fail in higher education, where students exercise substantial discretion in choosing courses and instructors. Finally, we propose a new identification strategy based on students' course histories and state the assumptions under which it yields unbiased estimates.

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

2.1 Data generating process

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

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

$$\nu_{ijct} = \mu_j + \epsilon_{ijct} \tag{2}$$

where X_{it} captures student i's background characteristics, C_{jct} are characteristics of other students taking class c with instructor j,[9] ρ_c is a course fixed effect, λ_t is a period fixed effect, and ν_{ijct} is a composite error term. The composite error is a combination of individual error ϵ_{ijct} and μ_j, which is instructor j's value-added to A^*_{ijct}.

Our estimands of interest are μ_j, instructor j's value-added, and the variance of the distribution of value-added, σ^2_μ, which characterizes the distribution of instructor quality. The standard deviation σ_μ is the average impact of having a one standard deviation higher value-added instructor. Our main simplifying assumption is that value-added is fixed across time t and across course c.[10] The key distinction between K-12 and higher education is that college students actively choose both their courses and their instructors. In K-12, centralized assignment systems limit the scope for selection, making lagged achievement a sufficient control. In higher education, students' choices reflect latent characteristics — career intentions, research interests, and private information about their own abilities — that lagged achievement cannot capture. The next subsection formalizes this concern and develops our approach to addressing it.

2.2 Identification under selection on observables

Identification of μ_j in the value-added literature relies on a selection-on-observables assumption: conditional on observable characteristics, assignment to instructors within a course is

[8] In elementary and secondary education, the conventional achievement measure is performance on a standardized test. Lacking standardized tests in higher education, we use 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-matriculation.

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

[10] Research in K-12 demonstrates that instructors improve over the course of their careers (Kini and Podolsky, 2016) and may have match-specific value-added that makes them better suited to teach subpopulations of students (e.g., Gershenson et al., 2022). These properties may hold in higher education as well. Though they are beyond the scope of this paper, they are important topics for future work.

as good as random. Formally, this requires

$$E[\epsilon_{ijct} \mid X_{it}, C_{jct}, \rho_c, \lambda_t] = 0. \tag{3}$$

The extent to which this assumption holds depends on whether the observable controls X_{it} capture the margins of selection that are also correlated with the outcome A^*.

In K-12 education, where class assignment is often centrally coordinated, research demonstrates that a parsimonious set of observable controls is often sufficient to satisfy Equation 3 (Kane and Staiger, 2008; Chetty et al., 2014a). The two most important controls are lagged student achievement (measured through prior standardized test scores) and class composition (measured through classroom averages of these lagged scores). Lagged test scores correlate with unobservable ability that is likely correlated with both future performance and instructor assignment; classroom averages serve as a sufficient statistic for selection patterns among classmates (Altonji and Mansfield, 2018). These controls work in K-12 largely because many students are assigned to classes by central allocation systems in which fairness, if not randomness, is seen as desirable. Existing studies that measure instructor value-added in higher education similarly rely on settings where course assignment occurs centrally, thereby restricting the channels through which students' characteristics can influence instructor assignment (e.g., Carrell and West, 2010; Figlio et al., 2015; DeVlieger et al., 2018). In most universities, however, students exercise substantial discretion in choosing both courses and instructors, and these choices may reflect unobservable characteristics that are correlated with outcomes. The next subsection formalizes this concern and develops our approach to addressing it.

2.3 Selection on unobservables in higher education

Suppose that the true data generating process for outcome A^* includes unobserved student characteristics Z_{it}:

$$A^*_{ijct} = X_{it}\beta + Z_{it}\omega + C_{jct}\gamma + \rho_c + \lambda_t + \nu_{ijct} \tag{4}$$

where Z_{it} represents latent characteristics of student i — such as career intentions, motivation, and latent abilities — that may influence both course and instructor selection and the outcome A^*. Estimation based on Equation 1, which omits Z_{it}, will produce biased estimates of instructor value-added whenever $\omega \neq 0$ and Z_{it} is correlated with instructor assignment.

As an 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 it to develop foundational knowledge for their research. The medical school type has an incentive to find the instructor who maximizes their likelihood of receiving a high grade. The chemist type has an incentive to find the instructor who provides the most chemistry-specific human capital.[11] Estimating instructor value-added to lifetime earnings without accounting for these different types will confound differences in earnings attributable to instructors with differences in earnings typical of doctors relative to chemists.

We propose that these unobserved types, while not directly observed, can be inferred from students' course selection behavior. Students reveal their types through the courses they choose: medical school types enroll in MCAT preparation courses in Psychology and Human Anatomy that chemist types would likely not take. We call this set of past and contemporaneous course choices a student's **course history**.[12] Conceptually, a student's course history serves as a rich proxy for the latent characteristics Z_{it} that drive selection into instructors' classes, just as lagged achievement serves as a proxy for selection in K-12. Though we cannot make unobservables observable, we can address selection by comparing students within the same course history "type," thereby restricting comparisons to students whose unobservables are plausibly similar.

The logic of this approach is familiar within education economics. In an influential methodological paper, Dale and Krueger (2002) proposed comparing the earnings of students who were admitted to the same set of colleges. In their setting, the unobservables were student aptitudes and preferences for various college attributes, as well as colleges' private observations of student qualities revealed through 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. Dale and Krueger argued that application portfolios could plausibly make the relevant unobservables as good as observable, since two students with the same portfolio might be so alike on unobservables that their actual college choices were effectively random.

Course histories play an analogous role to admissions portfolios. When students choose

[11]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. We demonstrate in Section 5 that forecast bias is indeed a problem.

[12]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.

courses, they act on many of their own unobservables: career goals, motivation, research interests, and social networks. In some respects, course histories are a noisier signal of student type than application portfolios, since course choices are also constrained by prerequisites, scheduling, and major requirements. However, these constraints cut both ways: they artificially steer students into courses based on university-wide factors unrelated to unobserved student characteristics, which may actually reduce selection bias. Moreover, course histories offer a benefit that application portfolios do not: we observe a *sequence* of choices that evolves as students progress through college, providing increasingly informative signals of student type over time.

Formally, we define a course history for student i in period t as the set of courses that student i has enrolled 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}:

$$A^*_{ijct} = X_{it}\beta + C_{jct}\gamma + H_{ict} + \rho_c + \lambda_t + \nu_{ijct}. \tag{5}$$

The fixed effect H_{ict} partitions students taking course c during period t into groups based on the similarity of their course histories. Section C describes the hierarchical clustering procedure used to construct these groups.

The validity of this approach rests on the following assumption:

Assumption 1 (Conditional Independence Given Course Histories)**.** *The following two conditions hold:*

1. ***Relevance****: Course history groups capture meaningful variation in latent student characteristics Z_{it}. Specifically, within a course history group, students are sufficiently similar on Z_{it} that the residual variation in Z_{it} does not generate omitted variable bias of practical importance.*

2. ***Conditional exogeneity****: Conditional on course history groups and observable controls, assignment to instructors is uncorrelated with potential outcomes:*

$$E[\epsilon_{ijct} \mid X_{it}, C_{jct}, H_{ict}, \rho_c, \lambda_t] = 0. \tag{6}$$

Assumption 1 would be violated if some unobserved characteristic — such as within-type variation in motivation, access to information, or social networks — is uncorrelated with a student's course history group but correlated with both instructor choice and outcomes. In other words, it requires that students who have taken similar courses up to and including the

current semester sort across instructors in course c in a manner that is effectively random with respect to their potential outcomes, perhaps due to registration timing, schedule constraints, or other idiosyncratic factors.

Course histories are unlikely to capture every dimension of student heterogeneity. For example, two students with identical course histories might differ in their access to peer networks, advising relationships, or private information about instructors gleaned from older students. If these residual differences are correlated with both instructor choice and outcomes, Assumption 1 would fail. The question is whether these within-group differences are large enough to generate economically meaningful bias — a question we take to the data in Section 5 .

3 Data and Setting

We use two sets of longitudinal administrative data on students: transcript data from Purdue University, and linked transcript-to-earnings data covering all public universities in Texas.

3.1 Purdue student panel

We use data from Purdue University, a selective public R1 institution in Indiana with a strong focus on STEM fields. Our dataset covers student transcripts from 2011 to 2023, providing detailed information on course enrollment, grades, pre-enrollment characteristics (such as entrance exam scores), and student demographics. The Purdue sample contains 1,530,649 student-course-semester observations across 3,569 unique courses taught by 8,625 instructors.

The Purdue data allow us to construct benchmark estimates of instructor value-added under conditional random assignment. In Fall 2018, Purdue instituted a "batch registration" procedure through which students submit ranked lists of course preferences and an algorithm assigns students to courses and sections.[13] This assignment mechanism generates conditional random assignment: students with the same or similar submitted preference lists have the same or similar probabilities of being assigned to any given instructor, which is true by construction.[14] We observe the preference lists for each student in every semester where students were assigned via algorithm: Fall 2018, Fall 2019, and every semester from Fall 2020 onward, including Spring semesters. A small number of students, such as athletes,

[13]The policy was initially adopted in response to an unusually large incoming cohort without a corresponding increase in course capacity (Mumford et al., 2024), but it has been maintained as Purdue's standard registration procedure in subsequent years.

[14]Notably, our analysis does not require the algorithm to be strategy-proof. We rely only on the fact that the algorithm uses submitted preferences, not true preferences, to determine assignment probabilities.

were exempt from algorithmic assignment and are excluded from our analysis. Appendix B contains more details about the algorithm.

This benchmark setting serves as the "gold standard" against which we evaluate our course history method and assess the extent of selection bias in conventional value-added estimation.

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 2023. The transcript data show every course taken by each student, including the instructor of record, course grade, and credits earned. The data also contain pre-enrollment characteristics (such as entrance exam scores) and student demographics. The Texas sample contains 36,993,297 student-course-semester observations across 97,516 courses taught by 116,094 instructors at 33 universities.[15]

These transcripts are linked to the state's unemployment insurance system, allowing us to track quarterly earnings for students who remain employed in Texas.[16] Our primary earnings measure is log quarterly earnings six-years after matriculation.[17] The dataset also includes instructor-level information, such as academic rank, demographic characteristics, and salary.

The Texas data offer three key advantages. First, we can estimate novel measures of instructor value-added to *earnings*, which is rarely possible in higher education and has not been done directly in K-12.[18] Second, we can explore the relationship between instructor characteristics and value-added. Third, the 33 Texas institutions span the full range of US public universities. On key observable dimensions — admissions rates, student-faculty ratios, and standardized test scores — the Texas universities are broadly representative of US public universities (Appendix Table A-1). The Texas sample also reflects typical patterns in the use of contingent instructors: 27% of instructors at the average Texas university are in

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

[16]Earnings for self-employed workers or those who move out of Texas are not observed. We observe earnings for 85% of students. We discuss implications of this attrition in Section 6.1.

[17]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.

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

contingent positions (e.g., instructor, lecturer), compared to 20% nationally. We revisit the role of contingent faculty in Section 7, where we consider counterfactual retention policies.

Finally, we collected student evaluations of instructors from online archives at four Texas universities: the University of Texas at Austin, Texas Tech, Stephen F. Austin State University, and the University of Texas at Tyler. Section 7 describes these data and their relationship to value-added in detail.

4 Estimation and implementation

This section describes how we estimate instructor value-added, implement the course history groups introduced in Section 2.3, and construct benchmark estimates using Purdue's conditional random assignment policy.

4.1 Value-added estimation

To estimate μ_j as specified in Equation 1, we begin by residualizing A^*_{ijct} on background and classroom characteristics. Under lagged achievement model, which uses only conventional K-12 controls, the residualized outcome is:

$$A^{LA}_{ijct} = A^*_{ijct} - \left(X_{it}\hat{\beta} + C_{jct}\hat{\gamma} + \hat{\rho}_c + \hat{\lambda}_t \right) \tag{7}$$

where X_{it} includes the student's previous-semester GPA, college entrance exam scores, academic level (freshman, sophomore, etc.), gender, race, and age; and C_{jct} includes classroom averages of lagged semester GPA and entrance exam scores.[19] The estimated coefficients and fixed effects come from a regression that includes μ_j in the residualizing step, ensuring that these coefficients are estimated using only within-instructor variation (Chetty et al., 2014a). Under the course history model, we augment the residualization with course history group fixed effects:

$$A^{CH}_{ijct} = A^*_{ijct} - \left(X_{it}\hat{\beta} + C_{jct}\hat{\gamma} + \hat{H}_{ict} + \hat{\rho}_c + \hat{\lambda}_t \right). \tag{8}$$

When Assumption 1 holds, A^{CH}_{ijct} is an unbiased estimate of $\mu_j + \epsilon_{ijct}$, and comparisons of outcomes across instructors within the same course history group identify instructor value-

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

added. We refer to estimates based on Equation 7 as *lagged achievement* estimates and those based on Equation 8 as *course history* estimates.

To estimate the variance of the value-added distribution σ^2_μ, we use a method-of-moments estimator similar to Ayllón et al. (2025). We split the sample of students in half within each course and semester, estimate instructor fixed effects separately on each subsample, and estimate σ^2_μ as the covariance between the two sets of estimates. This approach allows us to include all instructors who teach at least one course in the same semester as another instructor, without requiring a normality assumption on the distributions of value-added or errors.[20]

We construct three versions of value-added estimates, each suited to different applications:

1. **Empirical Bayes estimates** are used in all forecast bias tests. Because value-added estimates are subject to classical measurement error from large variances in ϵ_{ijct}, empirical Bayes shrinks noisy fixed-effects estimates of μ_j toward zero. Intuitively, it down-weights contributions from periods where an instructor has few students and attenuates all estimates more in subjects where σ^2_μ is small relative to σ^2_ϵ.[21]

2. **Jackknife empirical Bayes estimates** are used when regressing student outcomes on value-added, as in some forecast bias tests. If the same students' data were used to both estimate value-added and appear on the left-hand side, mechanical correlation would arise. Jackknife estimates address this by excluding the current period's data when predicting value-added for that period.[22]

3. **Raw average residuals** are used when value-added appears on the left-hand side of a regression. In this case, classical measurement error does not bias coefficients, since the measurement error is in the outcome variable rather than a regressor.[23]

For each application in subsequent sections, we note which version is used.

4.2 Implementing course history groups

We construct the course history similarity groups H_{ict} from Equation 5 using hierarchical clustering on the course histories of students enrolled in the same course across instructors.

[20] Our results are broadly similar under alternative variance estimators.

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

[22] For forecast bias tests which regress changes in average student outcomes on changes in value-added, we leave out both periods being differenced when constructing jackknifed estimates. Appendix C.3 contains more information about jackknife estimates.

[23] If unshrunk value-added estimates were used as regressors, their measurement error could correlate with the unobserved error in the model. When value-added is the outcome, this concern does not arise.

We encode each student's course history as an indicator vector over all courses offered at the institution, where an entry equals one if the student has taken that course in any period $t' \leq t$.

The clustering procedure requires three choices. First, we measure divergence between pairs of students using the Jaccard index, which calculates the fraction of shared courses relative to the total number of courses either student has taken. We use this measure because course history vectors are very sparse. Second, we measure divergence between groups using average linkage — the mean pairwise Jaccard divergence across all student pairs in each group. Third, we set the divergence threshold for defining final clusters at the mean of the levels at which each observation was first grouped. We chose this threshold *ex ante* to balance the tradeoff between within-group similarity and group size: lower thresholds produce more similar groups but leave more students ungrouped, while higher thresholds produce larger but more heterogeneous groups.[24]

After clustering, approximately 30% of student-course-period observations are in singleton groups. These tend to be students with unusual course histories relative to their classmates — for example, students taking a course outside their typical field or students with idiosyncratic schedules. In the main analysis, we exclude these students, as they have no within-group comparison to identify instructor effects. A robustness check in Appendix D.2 shows that pooling all singleton students within each course-period into a single reference group has little effect on results.

4.2.1 Illustrative examples

Hierarchical clustering based on course histories often yields student groups that align with intuitive understandings of how past coursework signals future intentions. We present three examples drawn from the Texas and Purdue data. In each case, all students in the example were enrolled in the same course during the same semester. Appendix Table A-4 lists the most common courses in each group's histories, along with average GPAs and test scores.

The first example identifies two groups among students taking Organic Chemistry at Texas A&M. One group's histories feature Engineering Mathematics, Computational Engineering, and Biomedical Engineering courses, a profile consistent with "chemist types." The other group commonly took Health core courses along with a Psychology and a Sociology course recommended for MCAT preparation, a profile consistent with "medical school types." The second example distinguishes "business types" (whose histories include Mar-

[24] See Appendix C.4 for additional implementation details. Appendix Table A-8 summarizes the resulting group structure: the mean number of groups per course-semester is approximately 11, with an average of 5 students per group.

keting, Accounting, and Management Information Systems) from "agricultural economics types" (whose histories include Agricultural Economics courses) among students taking Intermediate Microeconomics at Texas A&M. The third example separates "computer science types" from "engineering types" among students taking Calculus 3 at Purdue, based on whether their histories are dominated by Computer Science or Engineering core courses.

These groupings reveal meaningful unobservable heterogeneity among students in the same course. Importantly, this heterogeneity is correlated with instructor selection: Pearson's χ^2 tests reject independence between course history groups and instructor choice at the 5% level for Organic Chemistry and Calculus 3, and at the 10% level for Intermediate Microeconomics. Moreover, this differential selection cannot be addressed by lagged achievement alone. In both Intermediate Microeconomics and Calculus 3, the identified student types have nearly identical previous-semester GPAs, and the distributions of both GPA and entrance exam scores overlap substantially across groups (Appendix Figures A-1 and A-2). Controlling only for lagged achievement would therefore compare students across these distinct types, biasing value-added estimates.

4.3 Benchmark estimation at Purdue

Purdue's batch registration policy operates through the ranked lists of preferences submitted by students: students with identical preferences who want to enroll in an oversubscribed course are split randomly across sections by the algorithm.[25] To control for selection, we construct fixed effects for groups of students in the same course who submitted very similar preference lists, applying the same hierarchical clustering procedure described in Section 4.2 to a vector containing indicators for each student's six most preferred courses. By grouping on submitted preferences, we ensure that students in the same group have the same or very similar probabilities of being assigned to any given instructor.

The residualization for benchmark value-added augments the standard equation with assignment probability group fixed effects P_{ict}:

$$A_{ijct}^{CR} = A_{ijct}^{*} - \left(X_{it}\hat{\beta} + C_{jct}\hat{\gamma} + \hat{P}_{ict} + \hat{\rho}_c + \hat{\lambda}_t \right). \tag{9}$$

The fixed effect P_{ict} partitions students in course c during period t into groups with similar preference lists. Identifying variation comes from students in the same course with different instructors who had very similar probability distributions across sections. We refer to these

[25] Purdue uses a course registration algorithm developed by Müller et al. (2010). For more details, see Appendix B.

as *conditionally random* or *benchmark* estimates.

These benchmark estimates serve two purposes: they provide a "gold standard" comparison for our course history estimates, and they supply an additional source of identifying variation for forecast bias tests (Section 5).

4.4 Outcomes

We estimate value-added to two outcomes that capture different dimensions of instructor quality. The first is next-semester GPA, defined as the credit-weighted average of grades a student earns in the semester following the course of interest. Next-semester GPA is attractive because it is observed for nearly all students, it is available in real time for institutional decision-making, and it captures whether an instructor's impact persists beyond the classroom. However, GPA has well-known limitations: it is internal to the university, subject to grading norm differences across subjects and instructors, and reflects both the difficulty of future courses a student selects and the student's performance in those courses. The second outcome is log quarterly earnings six years after college entry, which we observe in the Texas data through linked unemployment insurance records. Earnings have a clearer connection to student welfare, are determined by the labor market rather than by institutional grading conventions, and are comparable across subjects and institutions. The primary limitation is that earnings are realized with a long lag, making them less useful for contemporaneous personnel decisions. Additionally, earnings are not observed for students who are self-employed, leave Texas, or are not in the labor force.

These two outcomes are intentionally complementary. GPA provides a near-term, high-frequency measure of academic performance; earnings provide a longer-term, market-based measure of human capital accumulation. Intermediate outcomes — such as major choice, graduation rates, or graduate school attendance — could further illuminate the channels through which instructors affect students, and we view these as promising directions for future work.

5 Validation: forecast bias

If Assumption 1 fails — that is, if unobserved factors influencing students' selection of instructors remain correlated with outcomes even after conditioning on course histories — then our value-added estimates will be biased. In this section, we define three tests for forecast bias that we use to evaluate our course history estimates, our lagged achievement estimates, and (at Purdue) our conditionally random benchmark estimates.

Following Chetty et al. (2014a), we define the forecast bias B of a value-added estimator

$\widehat{\mu_j}$ as $1-\alpha$, where α is estimated from the regression of residualized achievement A_{ijct} on $\hat{\mu}_j$ estimated on a separate sample:

$$A_{ijct} = \rho_c + \lambda_t + \alpha\widehat{\mu_j} + \psi_{ijct}. \tag{10}$$

If value-added estimates are unbiased, $\alpha = 1$: a one-unit increase in estimated value-added predicts a one-unit increase in student achievement out of sample. When $\alpha < 1$, value-added estimates systematically overpredict the differences across instructors, typically because unobservably higher-ability students sort to higher-quality instructors, inflating estimated differences. The true impact of moving to an instructor one standard deviation above the mean is not $\sigma_{\hat{\mu}}$ but $(1-B)\sigma_{\hat{\mu}}$.

We apply three tests that vary in their identifying assumptions, their statistical power, and their availability across our two settings. We present them in order from most to least general.

5.1 Test definitions

5.1.1 Individual test

The individual test is the simplest and most commonly applied diagnostic: it checks whether jackknife value-added estimates predict individual student outcomes out of sample (e.g., Petek and Pope, 2023). Let $\widehat{\mu_{jt}}$ be the empirical Bayes estimate of instructor j's value-added omitting all student-course observations from period t. The test regresses individual outcomes on these leave-out estimates:

$$A^*_{ijct} = \delta^{ind}\widehat{\mu_{jt}} + X_{it}\beta^{ind} + C_{jct}\gamma^{ind} + \rho_c^{ind} + \lambda_t^{ind} + \psi_{ijct}^{ind}, \tag{11}$$

where X_{it} and C_{jct} include the same controls as in the initial value-added estimation, and we additionally include the assignment group fixed effect P_{ict} when testing conditionally random value-added. An estimate of $\delta^{ind} = 1$ indicates no forecast bias.

The individual test is useful as a first-pass check: if value-added estimates cannot predict outcomes out of sample at all, something is fundamentally wrong with the estimation. However, this test is identified by the same selection-on-observables assumption that underlies the value-added estimation itself. Passing the individual test therefore does not independently validate the identifying assumption — it confirms internal consistency but not external validity. For this reason, we treat it as a necessary but not sufficient condition for credibility, and rely more heavily on the policy test for our validation.

5.1.2 Policy test

The policy test is our most cleanly identified test, but is available only at Purdue. It uses genuinely quasi-experimental variation from Purdue's conditional random assignment policy to evaluate value-added estimates constructed entirely from pre-policy, observational data.

The logic is as follows. We estimate value-added using only data from 2011-2017, before the batch registration policy was adopted. We then ask: do these observational estimates predict student outcomes during the post-2018 period, when randomization — rather than student choice — determined instructor assignment? If the observational estimates captured true instructor quality rather than selection, they should predict outcomes one-for-one even in this new regime. If they were contaminated by selection, the correlation will be attenuated, since the selection patterns that inflated the estimates are broken by randomization.

Let $\widehat{\mu_j}^{pre}$ denote empirical Bayes value-added estimated on pre-policy data. The test regresses individual student outcomes during the policy period on these estimates:

$$A_{ijct} = \rho_c + \lambda_t + \delta^{policy}\widehat{\mu_j}^{pre} + P_{ict} + \psi_{ijct}, \tag{12}$$

where P_{ict} are assignment probability group fixed effects from Equation 9, which control for the conditional random assignment. An estimate of $\delta^{policy} = 1$ indicates no forecast bias. This test is analogous to the exercise in Kane and Staiger (2008), which evaluated whether K-12 teachers' estimated value-added from observational data predicted their students' outcomes during a subsequent randomized experiment.

The policy test has strong internal validity but limited external validity: it can only be implemented at institutions with policies like Purdue's.

5.1.3 Prediction test

The prediction test assesses whether observable variables that were deliberately excluded from the value-added model would have been omitted variables — that is, whether they predict both value-added and outcomes. The intuition is straightforward: if our controls are sufficient to address selection, then additional observable characteristics (beyond those already in the model) should not predict our value-added estimates. If they do, it suggests that students are sorting on dimensions our model fails to capture.

We implement this test by checking whether students select instructors based on twice-lagged GPA, a variable excluded from our value-added estimation.[26] In Texas, where richer background data are available, we additionally check 9th-grade standardized test percentiles

[26] In K-12, it is common to use twice-lagged standardized test scores as the omitted predictor.

and parent income. To conduct the test, we first residualize both the omitted variable and the outcome (next-semester GPA) on the controls included in value-added estimation. We then regress the residualized outcome on the residualized omitted variable and form a predicted value $\hat{A}_{ijct}$. Finally, we regress this prediction on value-added:

$$\hat{A}_{ijct} = \gamma^{pred}\hat{\mu}_{jt}^{pred} + \psi_{ijct}^{pred}, \tag{13}$$

where $\hat{\mu}_{jt}^{pred}$ are jackknife empirical Bayes estimates constructed on the subsample of students with non-missing omitted variables. An estimate of γ^{pred} far from zero indicates that students are selecting on the omitted variable and that value-added estimates are biased.

This test is informative but has a specific limitation: it can only detect bias arising from selection on variables that are observable to the econometrician but excluded from the model. It has no power against selection on variables that are truly unobservable. Nevertheless, if value-added estimates pass this test for multiple excluded observables — particularly ones that are plausibly correlated with latent student characteristics, such as parent income — it provides indirect evidence that selection on unobservables is not severe.

5.1.4 Summary of tests and availability

Table 1 summarizes the three tests and their properties. All three tests are available at Purdue, where the policy test provides the strongest source of identification. In Texas, we apply the prediction test and individual test at each of the 33 universities. Across both settings and all tests, our criterion for a credible value-added estimator is that forecast bias coefficients should be close to 1.

Table 1. Summary of forecast bias tests

		Purdue	Texas
Test	**Identifying variation**		
Individual	Out-of-sample prediction	✓	✓
Prediction	Selection on excluded observables	✓	✓
Policy	Conditional random assignment	✓	
	Outcome		
	Next-semester GPA	✓	✓
	Income six-years post-matriculation		✓

The next two subsections present results from these tests, first at Purdue (Section 5.2) and then at Texas universities (Section 5.3).

5.2 Forecast bias results at Purdue

We begin with the policy test, our most cleanly identified test. The results, shown in Table 2, demonstrate that course histories reduce forecast bias dramatically relative to lagged achievement. The forecast bias coefficient for lagged achievement value-added is 0.482, far from 1, confirming that conventional controls are insufficient to account for unobservable student selection. In contrast, the coefficient for course history value-added is 0.855, much closer to 1 and indicating that course histories capture substantially more of the selection that contaminates lagged achievement estimates.

Table 2. Policy forecast bias test at Purdue

	Lagged Achievement	Course Histories
Value-added Estimate	0.531	0.675
	(0.165)	(0.108)
Course Preference Controls	X	X
Observations	147,139	142,817

Notes: The policy test regresses residualized next-semester GPA during the conditional random assignment period (2018-2023) on empirical Bayes value-added estimated from the pre-policy period (2011-2017). Assignment probability group fixed effects control for conditional random assignment. Standard errors are clustered at the period-subject level. A coefficient closer to 1 indicates less forecast bias.

The prediction test, shown in Table 3, examines whether students select instructors based on twice-lagged GPA, a variable excluded from value-added estimation. In the 2018-2023 sample (Panel A), all estimates are small and statistically indistinguishable from zero for all three methods. In the full sample (Panel B), lagged achievement value-added shows a small but statistically significant coefficient (0.037), while the course history estimate is essentially zero (-0.002). This test suggests that twice-lagged GPA is not a major source of omitted variable bias for any of our methods, though lagged achievement shows slightly more evidence of residual selection on this dimension.

Finally, the individual test, shown in Table 4, checks whether jackknife value-added estimates predict individual student outcomes out of sample. In both samples, lagged achievement value-added has coefficients far from 1 — 0.062 in the post-policy period and 0.171 in the full sample — indicating severe forecast bias. Course history value-added has coefficients near 1 (0.841 and 0.894), as does conditionally random value-added (0.813 in the post-policy period). While the individual test is identified by the same assumption as the initial estimation (Section 5.1.1), the magnitude of the gap between lagged achievement and course history performance is striking: lagged achievement estimates explain almost none of the out-of-sample variation in student outcomes.

Table 3. Prediction forecast bias test at Purdue

Panel A: 2018-2023	Lagged Achievement	Course Histories	Cond. Random
Value-added Estimate	0.004	-0.031	0.004
	(0.015)	(0.031)	(0.003)
Observations	165,519	154,506	155,482
Panel B: 2011-2023	Lagged Achievement	Course Histories	
Value-added Estimate	0.037	-0.002	
	(0.012)	(0.010)	
Observations	1,013,355	973,832	

Notes: The prediction test regresses the component of next-semester GPA predicted by twice-lagged GPA (after residualizing both on value-added controls) on jackknife empirical Bayes value-added. An estimate far from zero indicates selection on the omitted variable. Panel A uses the post-policy sample (2018-2023); Panel B uses the full sample (2011-2023). Standard errors are clustered at the period-subject level.

Table 4. Individual forecast bias test at Purdue

Panel A: 2018-2023	Lagged Achievement	Course Histories	Cond. Random
Value-added Estimate	0.062	0.841	0.813
	(0.052)	(0.080)	(0.077)
Observations	390,318	304,790	294,024
Panel B: 2011-2023	Lagged Achievement	Course Histories	
Value-added Estimate	0.171	0.894	
	(0.036)	(0.028)	
Observations	1,360,406	1,306,356	

Notes: The individual test regresses student-level residualized next-semester GPA on jackknife empirical Bayes value-added estimates. Panel A uses the post-policy sample (2018-2023); Panel B uses the full sample (2011-2023). Controls match those in the initial value-added estimation; conditionally random estimates additionally include assignment probability group fixed effects. Standard errors are clustered at the period-subject level. A coefficient closer to 1 indicates less forecast bias.

Across all three tests at Purdue, the results tell a consistent story. Course histories greatly reduce forecast bias relative to lagged achievement, and they perform comparably to the conditionally random benchmark that exploits Purdue's enrollment policy. Lagged achievement controls, which are typically sufficient in K-12, are clearly inadequate in the higher education setting.

5.3 Forecast bias results at Texas universities

We now assess whether course histories address selection bias at a broader set of institutions. Unlike Purdue, the 33 Texas universities have no policies restricting student course and instructor selection. In Texas, we estimate value-added to both next-semester GPA and log earnings six-years after matriculation. We apply the individual and prediction tests.

The individual test results in Texas mirror those at Purdue. Figure 1 shows point estimates and 95% confidence intervals for both methods at each university. For both value-added to GPA (Panel A) and value-added to earnings (Panel B), lagged achievement coefficients are significantly different from 1 at nearly every university, with more than half below 0.5 in magnitude. In contrast, nearly every course history coefficient is close to 1 and contains 1 in its confidence interval.

The prediction tests indicate little bias from selection on excluded observables. In Texas, we test three omitted predictors: twice-lagged GPA, 9th-grade standardized test percentiles, and parent income.[27] Figure 2 presents Bayesian averages of point estimates across universities with 95% prediction intervals. Neither method shows meaningful bias from selection on 9th-grade test scores or parent income, for either outcome. There is some indication of slight bias from selection on twice-lagged GPA for value-added to grades (average coefficient of 0.02 for lagged achievement and 0.01 for course histories), but the magnitudes are small. Value-added to earnings shows no bias from twice-lagged GPA under either method.

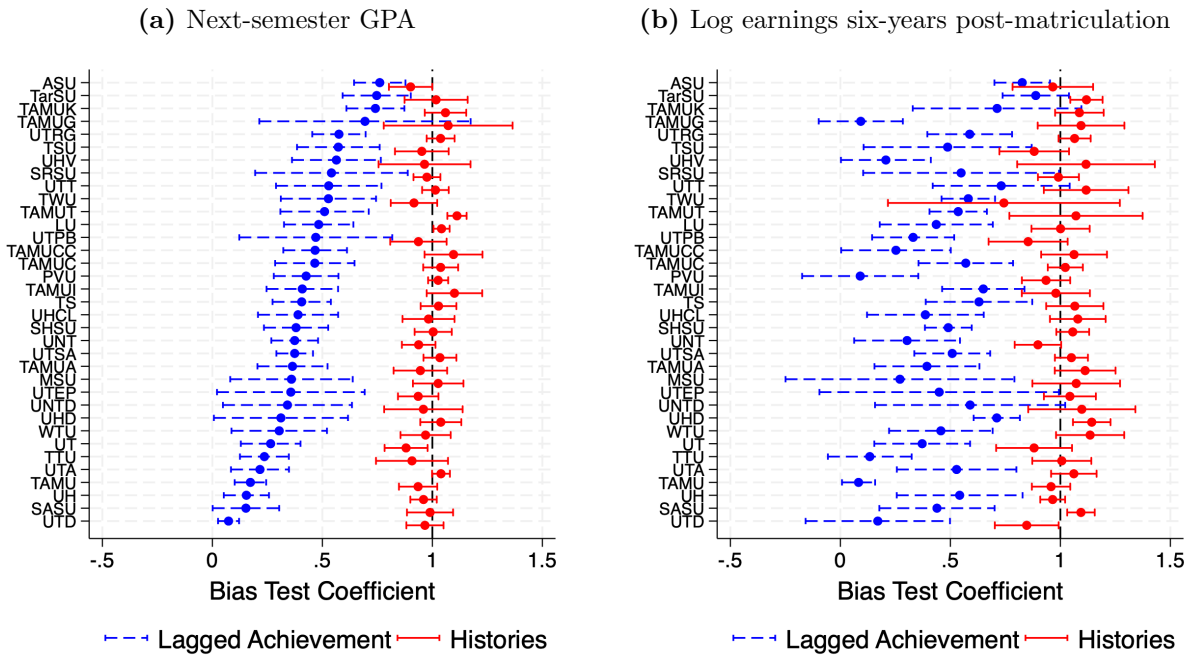
5.4 Summary

Across three forecast bias tests at Purdue and two tests at 33 Texas universities, the results are consistent: course histories substantially reduce forecast bias in value-added estimation relative to conventional lagged achievement controls.

At Purdue, lagged achievement value-added fails every test that does not share its identifying assumption, with coefficient of 0.48 in the policy test. Course history value-added performs far better, with coefficients between 0.82 on the policy test and comparably to the

[27]These variables are not available for all students, so we restrict to the subsample with available data when conducting this test.

Figure 1. Individual forecast bias test for Texas universities



Notes: The individual test regresses student-level residualized next-semester GPA (Panel A) and log earnings six-years post-matriculation (Panel B) on jackknife empirical Bayes value-added estimates. Controls match those in the initial value-added estimation. Each point is a separate university-level estimate. Standard errors are clustered at the period-subject level. A coefficient closer to 1 indicates less forecast bias.

conditionally random benchmark on the individual test. Estimates are in the range of validation results in the K-12 literature. In Texas, the same pattern holds across 33 universities and two outcomes: lagged achievement fails while course histories substantially reduce bias.

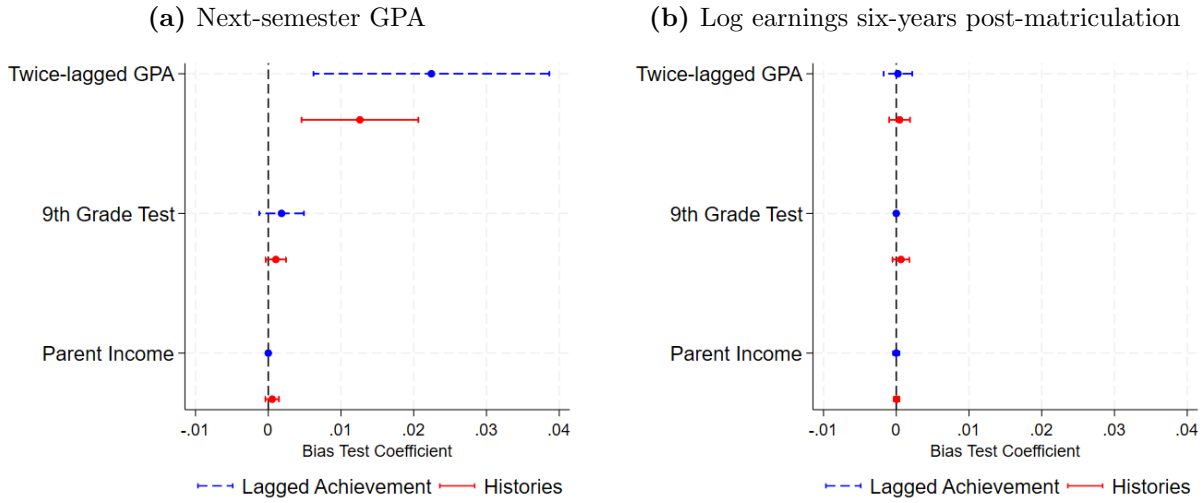
6 Effects of Post-secondary Instructors

Having validated our course history method, we now use it to document how much instructors matter for students' outcomes. We find substantial variation in instructor value-added to both grades and earnings, that observable instructor characteristics explain almost none of this variation, and that instructors appear to affect earnings largely through influencing students' choice of major.

6.1 Instructors impact students' future outcomes

Instructors vary meaningfully in their ability to impact students' outcomes. Table 5 summarizes the estimated variances of value-added to next-semester GPA and future earnings

Figure 2. Prediction forecast bias test for Texas universities



Notes: Each point is a Bayesian average across university-level estimates; bars show 95% prediction intervals. A coefficient far from zero indicates selection on the omitted predictor. Omitted predictors are twice-lagged GPA, 9th-grade Texas standardized test percentile, and parent income.

across Texas universities. Taking the square root of these variances gives the effect of moving from an average instructor to one with 1 standard deviation higher value-added within the same course.

An instructor with 1 s.d. higher value-added to GPA increases their students' next-semester GPA by 0.078 grade points, approximately 2.6% of the mean. For the average student, this roughly corresponds to raising one course grade from a B to a B+ the following semester.[28] Instructors also affect future earnings: an instructor with 1 s.d. higher value-added to earnings increases students' log earnings by 0.083 log points six years after college entry.[29]

These effects could operate through multiple channels. Instructors may vary in their ability to build human capital, or they may affect students' future course selections, major choices, or graduation probabilities. We investigate these channels in Section 6.3.

[28]This estimate is smaller than estimates of value-added for college algebra instructors on end-of-term exams in DeVlieger et al. (2018). A likely explanation is that DeVlieger et al. (2018) measure value-added to same-course grades, where the instructor's impact is more immediate, whereas we estimate value-added to grades in subsequent courses.

[29]This estimate includes only students with non-zero earnings in at least one quarter during the sixth year post-matriculation. We observe earnings for 85% of students. Attrition would be problematic for our analysis if instructors differentially affect students' likelihood of appearing in the earnings data. We assess the extent of this problem by estimating instructor value-added to the likelihood of appearing in the earnings data and find no meaningful variation.

Table 5. Variances of value-added distributions

	Next-semester GPA (1)	Log Earnings (2)
All	0.0061	0.0069
R1	0.0057	0.0059
Non-R1	0.0063	0.0072

Notes: Variances estimated within subject and institution using maximum likelihood estimation following Gilraine et al. (2020), then averaged across subjects and institutions, weighted by student-course-period observations. Universities are split by 2010 Carnegie classification: R1 universities have "very high research activity."

6.2 Observable characteristics do not predict value-added

If instructor value-added could be predicted from easily observed characteristics like rank, demographics, salary, then universities would not need to estimate it directly or collect proxies like student evaluations. We find that this is not the case. Table A-3 reports regressions of standardized value-added on a rich set of instructor characteristics in Texas: academic rank, race/ethnicity, gender, and nationality.[30] The R^2 values are 0.016 for value-added to GPA and 0.009 for value-added to earnings. Observable characteristics explain almost none of the variation in instructor quality.

This finding aligns with the K-12 literature, where observable teacher characteristics are similarly weak predictors of value-added.[31] The implication is that value-added represents genuinely new information — it cannot be inferred from CVs, hiring files, or salary schedules. Universities that wish to identify their most effective instructors must measure value-added directly.

Although the overall explanatory power is low, Table 6 reveals some heterogeneity in average value-added across characteristics. Coefficients represent standard deviation differences relative to the omitted category (assistant professors, white, male, native-born). In Panel A, there are no significant differences in value-added to GPA across academic ranks. However, for value-added to earnings, associate professors and contingent instructors have significantly higher value-added than assistant professors, and an F-test rejects equality of the rank coefficients.[32] This finding is relevant for the policy analysis in Section 7, where we

[30]Since our primary value-added estimates are fixed across time, we assign each instructor their modal observed academic rank.

[31]Seminal work by Hanushek (1971) and Ehrenberg and Brewer (1994) finds that K-12 instructor characteristics weakly predict student achievement. More recent work suggests that other factors, such as principal evaluations and certifications, have some predictive power (Jacob and Lefgren, 2004; Clotfelter et al., 2007).

[32]Previous work has found contingent instructors improve (Bettinger and Long, 2010; Figlio et al., 2015),

Table 6. Heterogeneity in value-added across instructor characteristics

	Value-added	
	GPA	Earnings
	(1)	(2)
Panel A: Academic Rank		
Full Professor	-0.007	0.007
	(0.016)	(0.017)
Associate Professor	0.004	0.020
	(0.016)	(0.017)
Contingent Instructors	0.006	0.029*
	(0.015)	(0.016)
Equality Test	0.730	0.285
N	62,325	59,066
Panel B: Race		
Asian	-0.028*	0.016
	(0.016)	(0.015)
Black	-0.029*	-0.018
	(0.015)	(0.017)
Hispanic	0.015	0.003
	(0.013)	(0.013)
Equality Test	0.091	0.362
N	62,388	59,122
Panel C: Female	0.001	-0.001
	(0.009)	(0.008)
N	62,388	59,122
Panel D: International	-0.053**	0.005
	(0.020)	(0.018)
N	62,388	59,122

Notes: Each column in each panel reports a separate regression of standardized value-added on the indicated characteristic, controlling for subject and institution fixed effects. The omitted category is assistant professor (Panel A), white (Panel B), male (Panel C), and native-born (Panel D). Observations are at the instructor level. Standard errors are clustered at the institution level. Coefficients represent standard deviation differences in value-added relative to the base category. F-test p-values in Panels A and B test equality of coefficients across categories.

consider retention policies targeted at contingent faculty.

6.3 Mechanisms behind instructor impacts

The variation in instructor effects on both grades and earnings raises a natural question: through what channels do instructors affect these outcomes? Instructors might improve students' human capital directly, or they might redirect students' academic trajectories toward different courses, different instructors, or different majors. We investigate these channels using Shorrocks-Shapley decompositions (Shorrocks et al., 2013), comparing how much of the outcome variance attributed to instructors is absorbed by future variables that proxy for specific channels. This analysis is descriptive, not causal. Because the future variables we condition on — major choice, course selection, graduation — are themselves endogenous to instructor effects, the decomposition cannot cleanly separate channels. In particular, if a high value-added instructor both builds human capital directly and redirects students toward higher-earning majors, controlling for major will absorb some of the direct human capital effect that happens to operate through students who also switch majors. The decomposition therefore provides an upper bound on the share of instructor effects attributable to sorting and a lower bound on the share attributable to direct human capital formation. With this caveat in mind, the magnitudes are nonetheless informative about the relative importance of different channels.

The decomposition compares attributions to instructor fixed effects across two specifications:

$$Y_{it} = \underbrace{X_{ijct}\beta}_{\text{Past Controls}} + \underbrace{\mu_j}_{\text{Instructor FE}} + \epsilon_{ijct} \tag{14}$$

$$Y_{it} = \underbrace{X_{ijct}\beta}_{\text{Past Controls}} + \underbrace{\mu_j}_{\text{Instructor FE}} + \underbrace{Z_{it}}_{\text{Future Variables}} + \eta_{ijct} \tag{15}$$

where course history fixed effects are included in past controls, and all decompositions are conducted within institution and subject. If adding a future variable Z_{it} substantially reduces the attribution to instructor fixed effects, this suggests that the instructor's impact operates partly through that channel. We decompose both next-semester GPA and earnings within each subject and institution, adding a variety of future variables. Table 7 reports results averaged across subject and institution.

Instructors affect future GPA through course and instructor choice. Panel A examines value-added to next-semester GPA. Without future controls, 4.3% of the variation in next-

reduce (Ehrenberg and Zhang, 2005), or have no effect on student achievement relative to tenure-track faculty.

Table 7. Shorrocks-Shapley decomposition of instructor effects

	Attribution	
	Instructors	Future Variables
	(1)	(2)
Panel A: GPA		
None	0.0405	.
Course Composition	0.0293	0.2849
Instructor Leniency	0.0345	0.1240
Subject Composition	0.0406	0.0016
All	0.0258	0.3150
Panel B: Earnings		
None	0.0856	.
Graduating Major	0.0499	0.4356
Bachelor's Attainment	0.0751	0.1440
Final Cumulative GPA	0.0696	0.2520
Persistence in Subject	0.0807	0.0316
All	0.0451	0.5078

Notes: Column (1) reports the share of outcome variance attributed to instructor fixed effects under each specification. Column (2) reports the share attributed to the indicated future variable(s). Each row adds the specified future variable(s) to the baseline specification in Equation 14. "None" reports the baseline attribution without future variables. All decompositions are conducted within institution and subject and averaged weighted by student-course observations.

semester GPA is attributed to instructors. Adding course composition — a measure of the difficulty of a student's future courses — reduces this to 3.1%, suggesting that about one quarter of the instructor effect on GPA operates through sending students to easier (or harder) courses.[33] Adding instructor leniency produces a similar reduction.[34] In contrast, controlling for the subjects of next-semester courses has almost no effect. When all three variables are included, the attribution to instructors drops to 2.7%, suggesting that nearly half of the instructor effect on GPA operates through future course and instructor choice.

Instructors affect future earnings primarily through major choice. Panel B examines

[33]We measure "course composition" as the average grade in the student's next-semester courses. The goal here is to distinguish higher GPA through human capital that affects future grades from instructor influence on students taking courses that give higher grades on average.

[34]We measure leniency for each instructor 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, then standardizing within each subject and semester. We then include the average leniency of a student's instructors next semester in the decomposition.

value-added to earnings. Without future controls, 8.5% of earnings variation is attributed to instructors. Adding graduating major reduces this to 5.0% — a 44% reduction — indicating that nearly half of the instructor effect on earnings operates through shifting students' choice of field. Bachelor's attainment and cumulative GPA also matter, but neither absorbs as much of the instructor effect as major choice. With all future variables included, attribution to instructors falls to 4.5%.

That major choice explains so much of the instructor effect on earnings is consistent with prior evidence on the returns to field of study. Cross-sectional variation in average earnings across majors spans a range larger than the college-high school earnings gap (Altonji et al., 2012). Andrews et al. (2016) find that business majors at Texas universities earn roughly 50% more than liberal arts majors within the same institution; Bleemer and Mehta (2022) estimates a 46% earnings increase for students barely above the GPA eligibility threshold for eligibility to major in Economics relative to their second-choice major; and Kirkeboen et al. (2016) find that earning a science degree instead of a humanities degree can triple early-career earnings. A back-of-the-envelope calculation illustrates the implied magnitudes: in a 25-student course, an instructor would need to redirect roughly 1-2 students from a lower-earning major to one paying 50% more to generate the effects our estimates and Shapley results suggest.

Even after controlling for graduating major, bachelor's attainment, and cumulative GPA, instructors still account for 4.5% of earnings variance. This residual attribution is a lower bound on the non-sorting channel, since the future variables we control for are themselves endogenous to instructor effects.[35] Nevertheless, the fact that substantial instructor effects persist after controlling for all observable downstream outcomes provides evidence that instructors matter for human capital formation above and beyond their influence on students' academic trajectories.

7 Value-added and Student Evaluations

We find that student evaluation scores are predictive of retention for contingent instructors at four Texas universities, especially for instructors in the bottom 5% of the evaluation distribution. Therefore, universities could improve student outcomes by making personnel decisions using value-added instead of student evaluations. We benchmark possible earnings gains using two counterfactual deselection exercises: one that replaces the bottom 5% of instructors, motivated by work in Hanushek (2009) and Chetty et al. (2014b), and one

[35]For example, if a high value-added instructor both improves human capital and induces students to choose higher-earning majors, controlling for major will absorb some of the human capital effect that operates through major choice.

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 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. Using data from four Texas universities, we investigate whether these subjective evaluations align with value-added measures of instructor quality. We scraped teaching evaluations during our sample period from the University of Texas at Austin, Texas Tech, Stephen F Austin State University, and the University of Texas at Dallas. We calculated average evaluation scores for each instructor over these periods and merged these averages with the corresponding value-added estimates.

These student evaluation surveys vary across institutions and years and incorporate a wide variety of questions. To create a standardized measure, we focus on three types of questions that appeared in all years and institutions. Questions in the "Instructor Score" and "Course Score" categories capture students' general impressions of the instructor and the course, respectively. Questions in the "Teaching Score" category specifically assess the instructor's effectiveness in teaching and conveying material. We standardize these average evaluation scores within each subject and institution to ensure comparability.[36]

The first two columns of Table 8 summarize bivariate regression estimates of the relationship between instructor value-added to GPA and earnings on these aggregated evaluation scores. Instructor value-added scores in each column are standardized within subject and institution to be on a common scale. The first column shows 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 third column of Table 8, 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

[36]For more details on question and categorization, see Appendix E.

Table 8. 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.

and the average grades given by other instructors teaching similar courses. The measures of leniency are then standardized within subject for comparability of estimates.

Instructors who are more lenient tend to receive higher ratings. For example, 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 with leniency is approximately twice the magnitude of the correlation with value-added. This provides suggestive evidence that students favor instructors who assign higher grades.[37]

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

7.2 Student evaluations and instructor retention

To show that student evaluations do enter personnel decisions in practice, 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

$$R_{j,s,y+1} = S_{j,s,y} + \lambda_s + \epsilon_{j,s,y} \tag{16}$$

$$R_{j,s,y+1} = \sum_{v=1}^{19} S^v_{j,s,y} + \lambda'_s + \epsilon'_{j,s,y}, \tag{17}$$

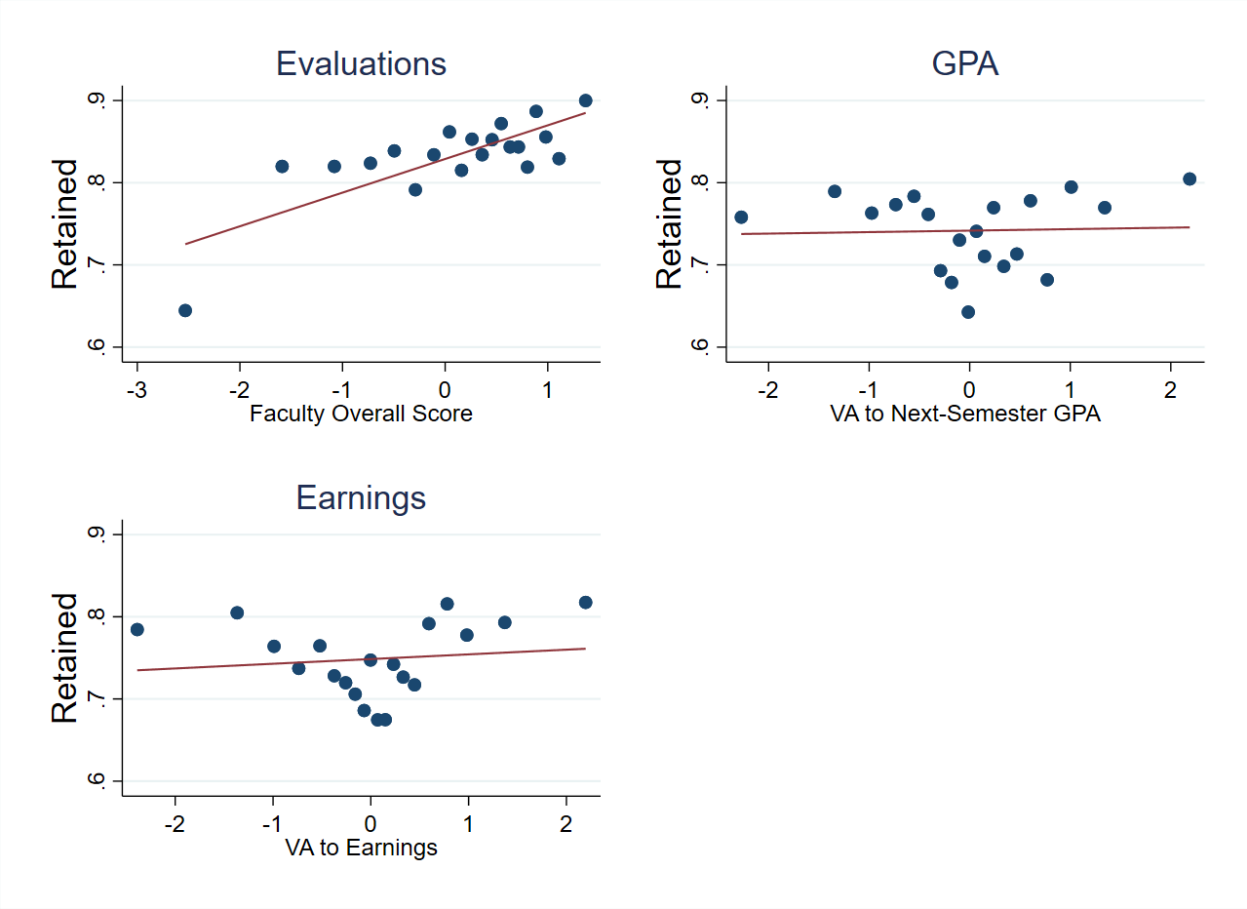
where $S^v_{j,s,y}$ is an indicator for instructor j having an evaluation score in the vth vigintile within subject s during year y, with the top vigintile as the reference group. We also estimate these equations with controls for estimates of value-added and vigintiles of estimated value-added, respectively.

We find that student evaluation scores predict retention. Appendix Table A-6 shows results from the linear regressions described in Equation 16, with and without value-added controls. Additionally, Figure 3 shows a binned scatter plot of retention and evaluations along with the estimated linear relationship. 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-7 shows results from the vigintile regression in Equation 17. 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 much less likely to be retained than the next lowest vigintile. The binned scatter plot in Figure 3 illustrates this nonlinearity. Above the bottom 5%, evaluations and retention have a somewhat linear relationship. However, the likelihood of retention plunges

measures.

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.

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.

7.3 Counterfactual deselection exercise

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

We calculate the average per-student increase in earnings six-years post-matriculation 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 0.083 log points; 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. For a student who has a bottom 5 percentile instructor, the average increase in earnings six-years post-matriculation from replacing an instructor in the bottom 5% is

$$G = 2.063 \times \$8,638 \times 0.083 = \mathbf{\$}1,479,$$

which is about 18% of median student earnings. These gains are substantial.[38]

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 possible groups of instructors with an average 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 9 reports results from this exercise. Both income and gains in income from replacement are highly skewed distributions. To limit the effects of outliers, 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. The average gain for students is \$300.77, which is about a 3% average quarterly earnings in our sample. Out of any single group, we find that the largest gains are from replacing contingent instructors. The average gain from replacing all bottom 5% contingent instructors is \$168.02, which is nearly 2% of the mean. Replacing instructors in the bottom 5% of assistant, associate, and full professors would result in smaller gains.

In practice, the university's retention policies are constrained. Many instructors cannot be dismissed due to tenure or other long-term contracts. Even among contingent instructors

[38]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.

Table 9. Average gains for deselection of the bottom 5%

	Average Gain	Percent Gain over Average Earnings
Replace All	\$300.77	3.33%
Replace Contingents	\$168.02	1.86%
Replace Assistants	\$36.67	0.040%
Replace Associates	\$51.27	0.057%
Replace Full Professors	\$45.24	0.50%
Replace Tenure Track	\$96.36	1.07%

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 mean of \$9041. Means were constructed by averaging winsorized gains for students with non-zero gains, then multiplying by the fraction of students with non-zero gains.

with low student evaluations, the likelihood of retention is still greater than 50%. It may not be feasible — or even desirable — to replace 5% of instructors annually. Within these constraints, this section explores the potential gains from alternative programs that maintain the current number of annual replacements but bases retention decisions on instructors' value-added rather than student evaluations.

In order to incorporate the university's constraints, we use the empirical relationship between student evaluations and retention observed for contingent instructors at Texas Tech. Implicitly, we assume that this relationship reflects the institution's optimal decisions when considering hiring, training and development costs that are important to the retention decision. Furthermore, this empirical relationship incorporates the natural turnover rate of contingent instructors.

We show the earnings gains from a policy based on value-added to earnings. While the policy based on value-added to earnings is likely not implementable, due to the long measurement horizon, the counterfactual gains give a first-best estimate of possible gains to earnings from any type of value-added based policy. Institutions could use a value-added measure that is correlated with value-added to earnings to achieve gains in earnings.

Finally, decision-makers would need to account for the fact that value-added estimates are noisy measures of true instructor quality. We show feasible policy gains that treat value-added as estimates by scaling instructor value-added by reliability, which largely depends upon the number of students used to estimate an instructor's value-added. Additionally, we show returns from the infeasible policy that treats value-added estimates as true value-added to estimate an upper bound on possible gains.

To implement the realistic counterfactual, we begin by predicting the probability that a contingent instructor j would be retained in year y. We construct this probability in

two ways: first, by using the prediction given from our estimates of Equation 17 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 across year and subject, the regressions and predictions included year and subject fixed effects. For instructors that cannot be replaced by the policy (tenure-track instructors and first-time instructors) we assign both retention probabilities as 1.

The model gives a probability that an instructor is retained under the status quo, which uses student evaluations for retention, and under the counterfactual which uses value-added. We estimate average earnings changes using the predicted retention probabilities for the full range of the value-added distributions, as well as a policy that only uses these retention probabilities for replacing the bottom 5% of instructors.

Given these retention probabilities, instructor j's value-added standardized within subject s, $\mu^z_{j,s}$, is the subject-specific variance of value-added to earnings $\sigma^2_{\mu,s}$, and student i's earnings six-years post-matriculation Y_i, the predicted gains $G_i(a)$ for student i under replacement regime a are:[39]

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

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 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.[40] 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 10 shows average infeasible and feasible gains for these policies. Columns 1 and 2 show changes from the "status quo" policy. In fact, 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) would be *reduce* average earnings. However, because the status quo is an approximation of the true policy that the university uses, these changes are small in magnitude — 0.2% of the mean. Applying the evaluation-based retention policy

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

[40] Appendix C.2 gives the exact formula.

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%.[41]

On the other hand, replacing instructors based on value-added to earnings 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 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-matriculation, 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 in the production of human capital during students' post-secondary educations. However, estimates of post-secondary instructor impacts have, thus far, been limited to a small set of unique institutions with uncommon enrollment policies.

[41] 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.

Table 10. 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.

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 an observational 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 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 0.078 points and increase their income by 0.083 log points. Furthermore, we document that student evaluations and instructor value-added to earnings six-years post-matriculation are essentially uncorrelated, highlighting a shortcoming of the use of evaluations in making personnel decisions. We find that a policy which uses value-added to earnings to deselect contingent instructors 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 on instructor effects averaged across all students. Future research could investigate heterogeneous instructor effects by race, gender, or ability.[42]

We view our two main 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 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 also estimated value-added to a variety of other outcomes. Using these outcomes, we found suggestive evidence that value-added to GPA maybe operating through affecting student's course choices next semester, and that value-added to earnings may be operating through student's major choice.

We expect that our method can be applied more systematically to other outcomes of interest in higher education that lie between GPA and earnings.[43] These more intermediate outcomes could both provide institutions with metrics of instructor quality that both are correlated with earnings and give deeper insight into the mechanisms behind the instructor impacts we document in this paper.

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

[43] 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

John M Abowd, Francis Kramarz, and David N Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 1999.

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.

Joseph G Altonji, Erica Blom, and Costas Meghir. Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annu. Rev. Econ.*, 4(1): 185–223, 2012.

Rodney J. Andrews, Jing Li, and Michael F. Lovenheim. Quantile treatment effects of college quality on earnings. *Journal of Human Resources*, 51(1):200–238, 2016. ISSN 0022-166X. doi: 10.3368/jhr.51.1.200. URL https://jhr.uwpress.org/content/51/1/200.

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.

Sara Ayllón, Lars J Lefgren, Richard W Patterson, Olga B Stoddard, and Nicolás Urdaneta Andrade. 'sorting'out gender discrimination and disadvantage: Evidence from student evaluations of teaching. Technical report, National Bureau of Economic Research, 2025.

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.

Zachary Bleemer and Aashish Mehta. Will studying economics make you rich? a regression discontinuity analysis of the returns to college major. *American Economic Journal: Applied Economics*, 14(2):1–22, 2022.

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.

Joseph J Doyle Jr, Steven M Ewer, and Todd H Wagner. Returns to physician human capital: Evidence from patients randomized to physician teams. *Journal of health economics*, 29 (6):866–882, 2010.

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.

Seth Gershenson, Cassandra M. D. Hart, Joshua Hyman, Constance A. Lindsay, and Nicholas W. Papageorge. The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy*, 14(4):300–342, November 2022. doi: 10.1257/pol.20190573. URL https://www.aeaweb.org/articles?id=10.1257/pol.20190573.

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.

Tara Kini and Anne Podolsky. Does teaching experience increase teacher effectiveness? a review of the research. *Learning Policy Institute*, 2016.

Lars J. Kirkeboen, Edwin Leuven, and Magne Mogstad. Field of study, earnings, and self-selection*. *The Quarterly Journal of Economics*, 131(3):1057–1111, 05 2016. ISSN 0033-5533. doi: 10.1093/qje/qjw019. URL https://doi.org/10.1093/qje/qjw019.

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. 2024.

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.

Lena Shi. Do college faculty affect students' long-term achievement? evidence from developmental courses. 2025.

Anthony F Shorrocks et al. Decomposition procedures for distributional analysis: a unified framework based on the shapley value. *Journal of Economic Inequality*, 11(1):99–126, 2013.

A Appendix Tables and Figures

Table A-1. Comparison of institution characteristics.

	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-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."

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

	VA to GPA	VA to Earnings
	(1)	(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)
Contingent Instructors	-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

A: Organic Chemistry Chemistry	Chemist Type (1)	Medical School Type (2)
Commonly Taken Courses	BMEN101 BMEN253 CHEM119 ENGR102 MATH151 VTPP434	CHEM120 HLTH210 HLTH236 HLTH240 PSYC107 SOCI205
Last Semester GPA	3.33	3.67
Admissions Test Percentile	86	59
Selection Test p-Value	0.04	
B: Intermediate Micro	Business Type	Agricultural Econ Type
Commonly Taken Courses	ACCT229 ACCT230 BUSN101 ECON202 ISTM210	AGEC117 AGEC217 AGEC314 AGEC340 POLS207
Last Semester GPA	3.30	3.32
Admissions Test Percentile	76	62
Selection Test p-Value	0.06	
C: Calculus 3	Computer Science Type	Engineering Type
Commonly Taken Courses	CS18000 CS18200 CS19100 CS19300 MA16200	ENGR16100 ENGR16200 HONR19901 HONR19902 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 χ^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. Rank-rank correlation between value-added measures

VA Outcome	GPA	Earnings	Bachelor's	Observed Earnings	Average Course Grade	Average Major Income	Time To Degree	Earnings 10 Years Post
GPA	1.000	-	-	-	-	-	-	-
Earnings	0.540	1.000	-	-	-	-	-	-
Bachelor's	0.640	0.521	1.000	-	-	-	-	-
Observed Earnings	0.514	0.525	0.569	1.000	-	-	-	-
Average Course Grade	0.582	0.522	0.590	0.715	1.000	-	-	-
Average Major Income	0.515	0.602	0.502	0.543	0.540	1.000	-	-
Time to Degree	0.463	0.447	0.515	0.535	0.532	0.503	1.000	-
Earnings 10 Years Post	0.509	0.586	0.487	0.507	0.504	0.529	0.480	1.000

Notes: This table shows rank-rank correlations between instructor ranks in several value-added measures. The outcomes used for each different measure, in order, are next-semester GPA, earnings six-years post-matriculation, Bachelor's attainment, having any earnings observed six-years post entry, average grade of a student's classmates next semester, average income of a student's graduating major, time to degree, and earnings ten-years post-matriculation. Ranks are constructed within institution and subject.

Table A-6. 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.

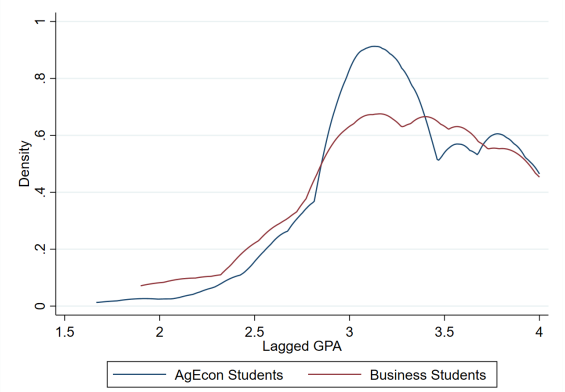
Table A-7. Non-linear retention of contingent instructors

	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

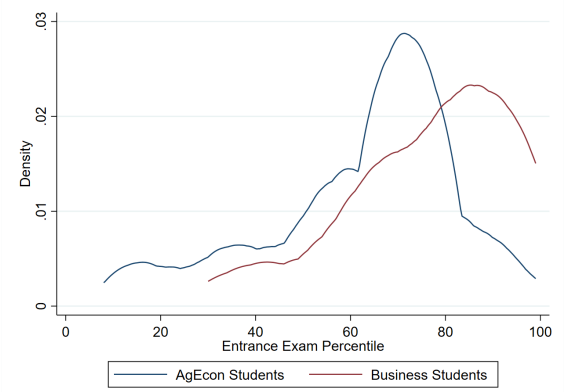
Notes: These regressions are similar to those in Table A-6, but use vigintiles of instructor scores from student evaluations. Estimates are relative to the top vigintile of instructor scores. Column 2 controls for vigintiles of value-added to GPA and earnings.

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

(a) Lagged GPA



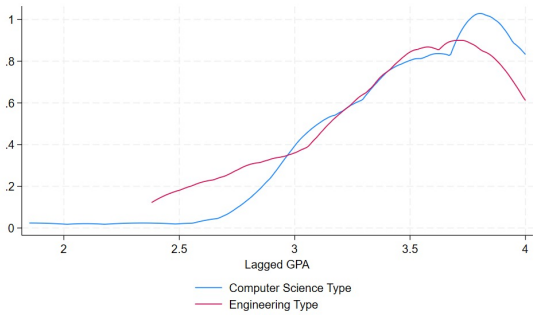
(b) Entrance Exam Percentile



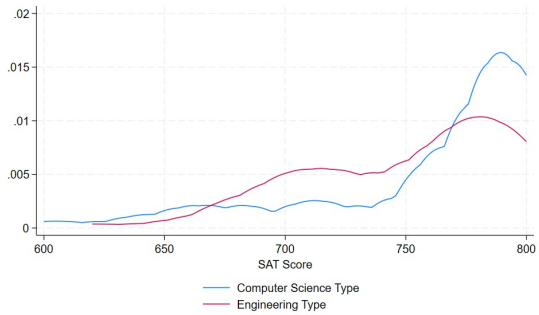
Notes: This figure shows smoothed kernel density plots for students in two different course history groups taking Intermediate Micro in the same semester at Texas A&M.

Figure A-2. Lagged achievement for computer science and engineering types in calculus 3 at Purdue

(a) Lagged GPA

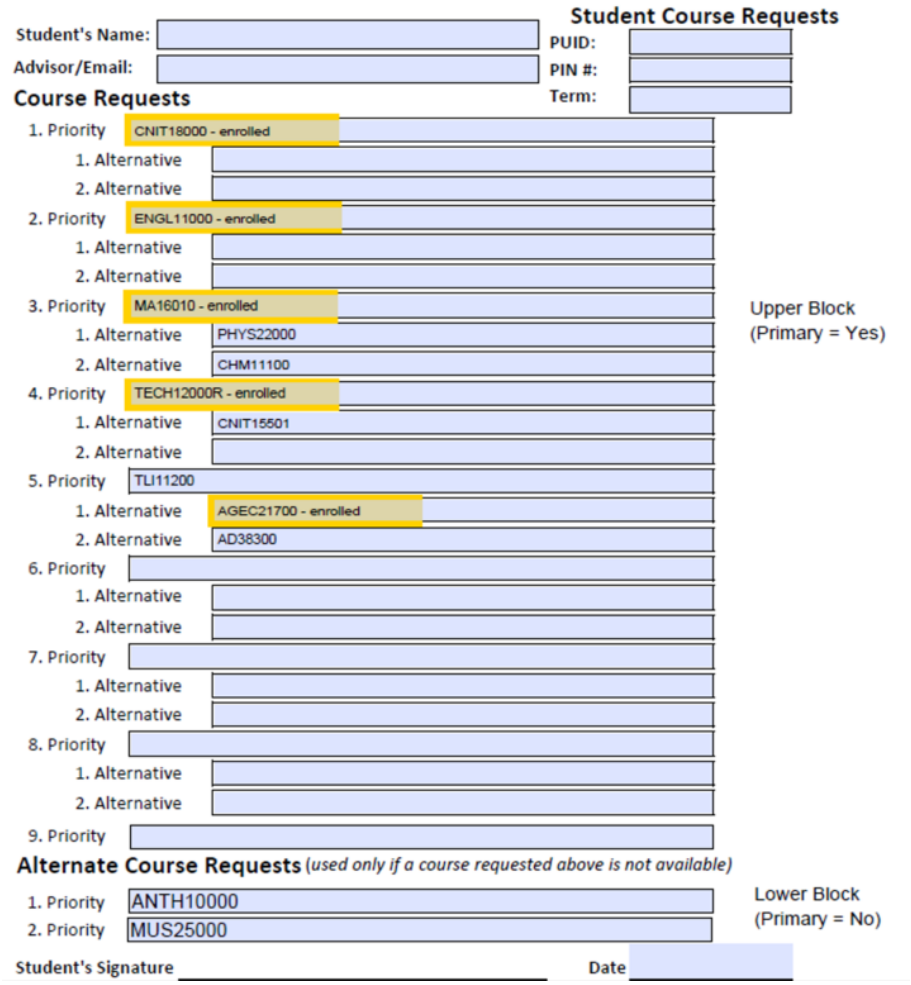


(b) SAT Score



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:

$$weight(a \in dom(R)) = 0.9^{prior(R)} \times 0.5^{alt(a)}, \tag{18}$$

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

In order to estimate the variance of value-added, we use a split-sample, method of moments estimator, inspired by Ayllón et al. (2025). Relative to existing estimates, this allows us to include instructors who are only in the data for one period, and also allows us to avoid making strong assumptions of normality on the distribution of value-added or the distribution of individual error.

To estimate the variance of the value-added distribution to outcome A_{ijct}, we begin by randomly splitting the sample in half, within each course-instructor semester. We denote the two subsamples with superscripts 0 and 1. We then estimate the fixed effects regression

$$A^*_{ijct} = X_{it}\beta^s + C_{jct}\gamma + \rho^s_c + \lambda^s_t + \mu^s_j + \epsilon_{ijct}, \tag{19}$$

including additional fixed effects H_{ict} and P_{ict} for course history value-added and conditionally random value-added respectively. This gives us unshrunk fixed effects estimates $\hat{\mu}^0_j$ and $\hat{\mu}^1_s$ for each instructor. We then estimate the variance of value-added as the covariance between these two estimates:

$$\hat{\sigma}^2 = \widehat{Cov}(\hat{\mu}^0_j, \hat{\mu}^1_j)$$

Similar to Rose et al. (2022), when the true variance of value-added is very close to zero, the estimated covariance can be zero. To get our final variance estimate, we take $\max(\hat{\sigma}^2, 0)$. This introduces a small degree of bias into our estimates, but is similar in principle to a shape restriction from a maximum likelihood estimator.

C.2 Empirical bayes estimation

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

To estimate empirical Bayes estimates of value-added, we begin by residualizing the outcomes A^*_{ijct} as in Equation **??** and obtain residuals A_{ijct}. We then estimate σ^2_μ and σ^2_ϵ 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}:

$$\begin{aligned} m_j &= \sum_t w_{jt} \overline{A}_{jt} \\ w_{jt} &= \frac{h_{jt}}{\sum_t h_{jt}} \\ h_{jt} &= \frac{1}{\hat{\sigma}^2_\mu + \frac{\hat{\sigma}^2_\epsilon}{N_{jt}}} \end{aligned}$$

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 **??**.

C.3 Jackknife empirical bayes estimates

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

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

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

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 $\mathbf{H_{ict}}$[44], 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 $i \neq j$:

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

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.

[44]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-8 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.

Table A-8. 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 ≥ 15	0.505	.

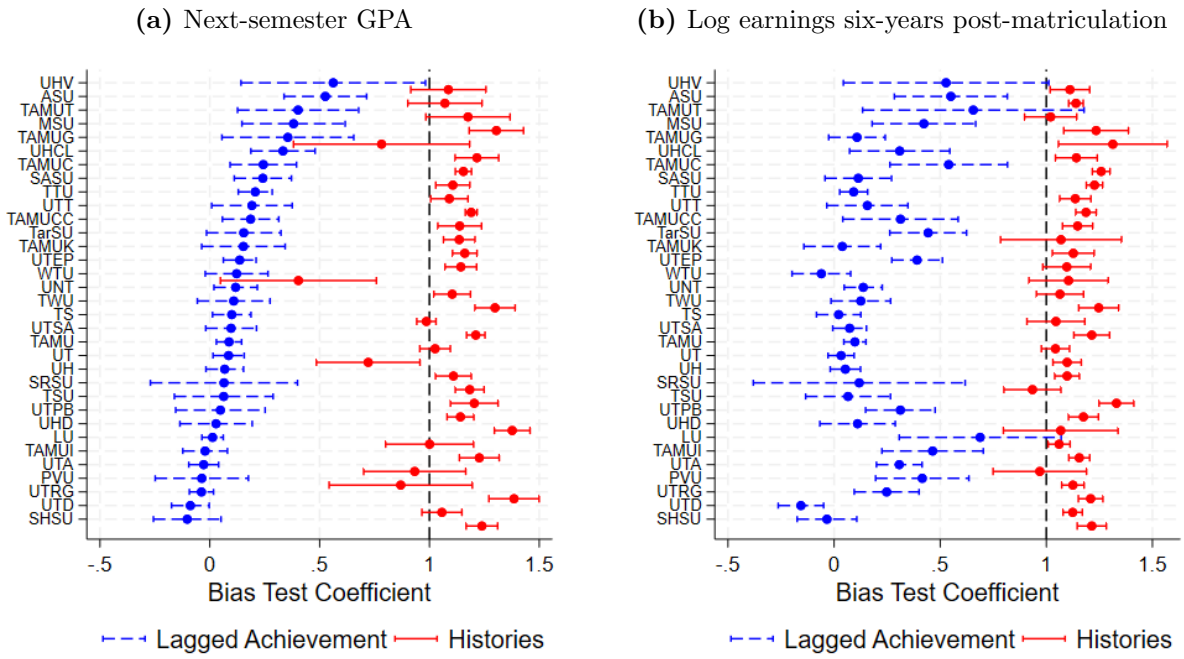
Notes: This table shows statistics on the course history similarity groups for our Texas sample. The units of observation for rows 1 and 2 were course-periods. The units of observation for rows 3 and 4 were student-course-periods.

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.

Figure A-4. Course teaching roster changes forecast bias test for Texas universities



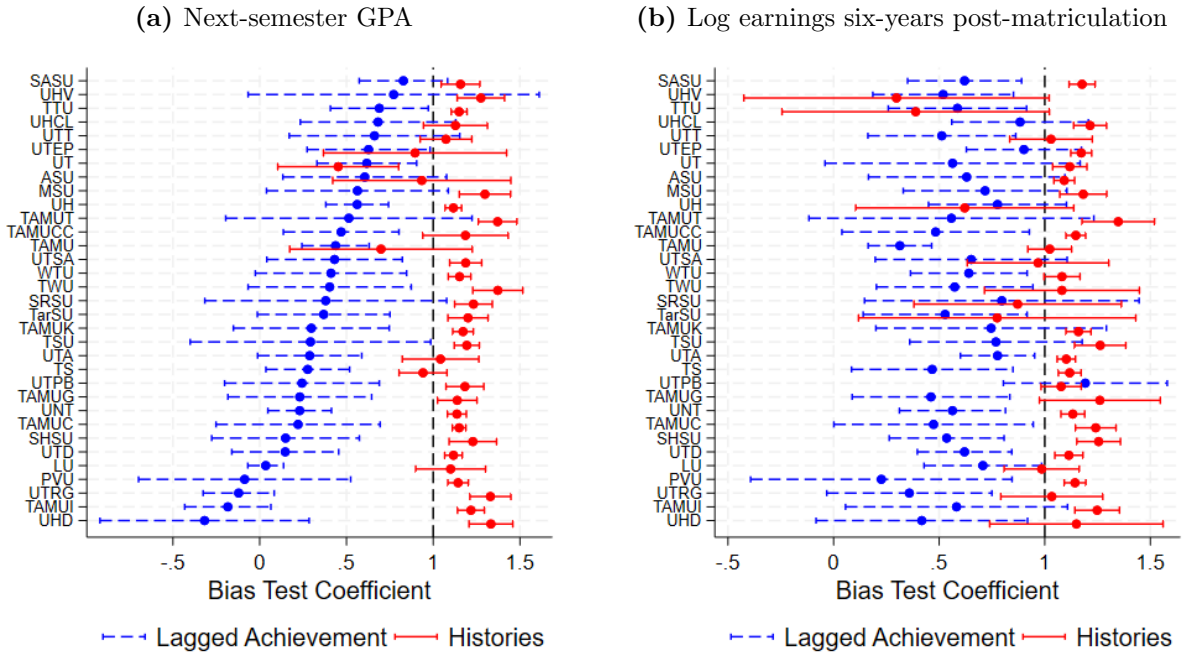
Notes: The teaching roster changes forecast bias test leverages year-to-year variation in teaching assignments to assess whether changes in residual student achievement are predicted by shifts in instructor value-added, with estimates regressing students' residualized next-semester GPA on changes in 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.

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 **??**.

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-matriculation that include ungrouped students. Similar to estimates using only grouped students, estimates with ungrouped students substantially reduce forecast bias relative to estimates that use only lagged achievement to control for student selection.

Figure A-5. 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 A-9 shows average variances across Texas institutions 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 **??** 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

Table A-9. Variances of value-added distributions with ungrouped students

	σ^2_μ
Next Semester GPA	0.014
Log Earnings	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.

instructor. For example, if students could perfectly predict teaching rosters and waited to take an intermediate-level economics course from a particular instructor, this test would be biased. Therefore, we rely on the assumption that most changes in teaching rosters are unexpected by students, or that students do not react to these changes.

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

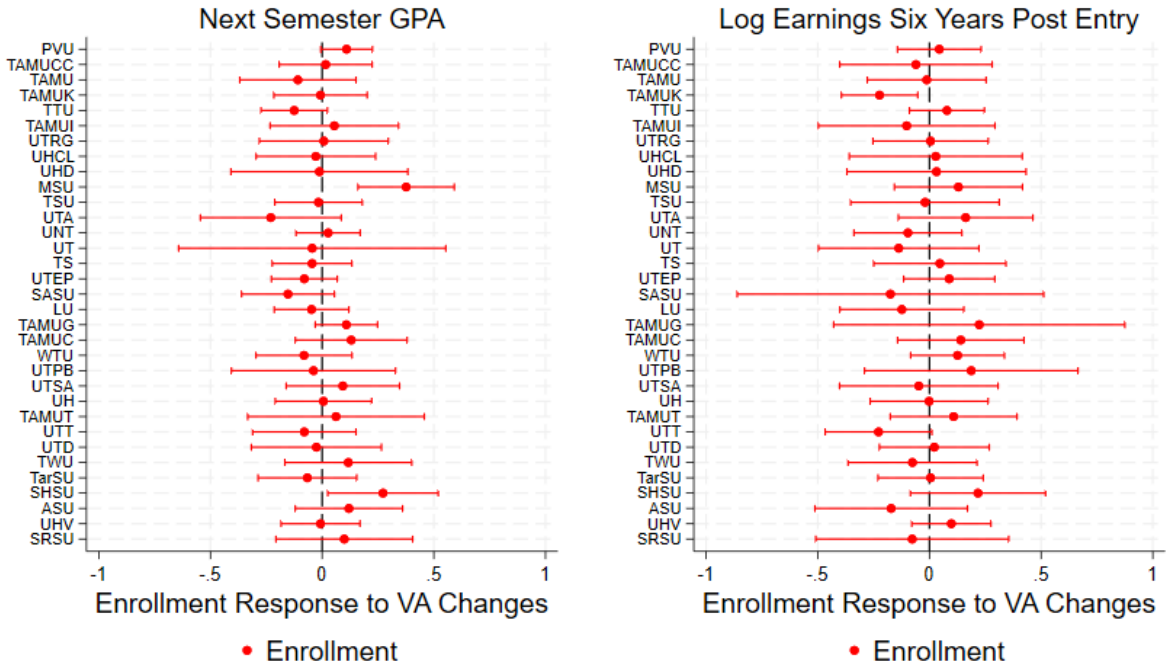
$$\Delta N_{slt} = \delta \Delta M_{slt} + \xi'_{slt}$$

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

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 enrollment 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.

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.

E Student Evaluation Questions

Table A-10 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-11 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-10. Evaluation Question Categorization

Category	Question Text
Instructor Score	Overall this instructor was effective Overall, the instructor was an effective teacher
Teaching Score	The course objectives were specified and followed by the instructor The instructor demonstrated knowledge of the subject The instructor stimulated student learning The instructor presented the information clearly The instructor emphasized the major points and concepts
Course Score	Overall this course was a valuable learning experience Overall, this course was a valuable learning experience
Soft Skills Score	The instructor treated all students fairly The instructor treated all students with respect The instructor welcomed and encouraged questions and comments 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 The testing and evaluation procedures were fair 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-11. 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.