Instructor Value-Added in Post-Secondary Education*

Merrill Warnick†
Jacob Light‡
Anthony Yim§

April 20, 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 accounts 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 0.05 points, and earnings six-years post-matriculation by about 5%. 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 as much as a 4% increase in for affected students.

*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 conclusions of this research do not necessarily reflect the opinion or official position of the Texas Education Research Center, the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, or the State of Texas. All opinions expressed in this manuscript are those of the authors and do not represent the opinions of Purdue University. 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 distort incentives in the classroom and cause grade inflation (Nelson and Lynch, 1984; Eiszler, 2002) and reflect systematic biases related to instructor gender,

[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).

race, and other attributes (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 course may 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 may 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 identifying value-added using only comparisons within these groups, 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. At Purdue University, an unusual batch registration policy assigns students to courses conditional on submitted preferences, generating quasi-random instructor assignment among students with similar preferences. 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.

[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.

To further evaluate the performance of our new estimator, we conduct a series of forecast bias tests that assess value-added estimate stability in out-of-sample predictions. 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 a benchmark of 1. We conduct three forecast bias tests, two which test out-of-sample predictive accuracy, and one which tests for selection on sets of student observables not included in value-added estimation. Our strongest test leverages conditional randomization coming from Purdue's assignment policy, using value-added estimated with data from before Purdue's policy to predict student outcomes during the later period, when the policy was in place.

While our value-added estimates using course histories perform well in these validation tests, we find that estimates including only conventional value-added controls, similar to the lagged achievement model standard in K-12, perform poorly in this setting. When evaluated against Purdue's randomization benchmark, the lagged achievement estimator yields a forecast bias coefficient of 0.282. Our course history estimator, by contrast, yields a coefficient of 0.861. Similar findings emerge from forecast bias tests using other sources of variation, where our course history estimates perform nearly as well as an estimator that exploits the randomization policy directly. Furthermore, the within-subject rank correlation between course history and benchmark conditionally random value-added estimates is 0.78. The results confirm that selection is a first-order problem in higher education that standard K-12 controls cannot solve, and that course histories substantially ameliorate this problem.

We next show that the method generalizes by estimating value-added at all 35 public four-year universities in Texas. These institutions enroll nearly 10% of all four-year post-secondary students in the US, spanning large flagship R1 institutions to smaller regional schools. The Texas data allow us to estimate value-added to earnings, linked through unemployment insurance records — making ours one of the first papers to estimate instructor value-added to earnings directly, rather than through the mediating channel of test scores. Forecast bias tests across all 35 universities confirm that course history controls address selection in this broader setting: for both value-added to grades and earnings, nearly all bias coefficients are close to 1. The lagged achievement model performs poorly in nearly all cases; more than half of these forecast bias coefficient estimates are less than 0.6.

We use our course history method to develop novel insights about the importance of instructors in higher education. Our estimates document that individual instructors matter substantially for students' outcomes. An instructor with one standard deviation higher value-added increases students' next-semester GPA by 0.05 grade points and earnings six-years post-matriculation by 0.05 log points. Variance decompositions suggest that future course

selection is a primary channel through which instructors influence next-semester grades, while instructors' differential influence on students' majors account for about half of the variation in value-added to earnings. Strikingly, observable characteristics (e.g., rank, demographics, salary) explain almost none of the variation in value-added, with R^2 less than 0.02 in regressions of value-added on a rich set of instructor observable characteristics. Value-added, therefore, is genuinely new information to the institution not available from CVs or resumes.

Finally, we show that the disconnect between evaluations and value-added has direct consequences for personnel policy. Using students' subjective evaluations of instructors at a subset of the Texas universities, we first document that student evaluations are more strongly correlated with instructor grading leniency than with value-added to GPA, and are essentially uncorrelated with value-added to earnings. However, evaluations strongly predict retention, particularly among contingent instructors. A counterfactual policy in the style of Hanushek (2009) that replaces evaluation-based deselection of the bottom 5% of contingent instructors with value-added-based deselection would increase quarterly earnings of affected students by approximately 4%, demonstrating that universities have meaningful room to improve student outcomes by reorienting personnel decisions toward objective measures of instructor quality.

This project makes three primary contributions. First, and most centrally, we develop and validate a scalable method for estimating instructor value-added in higher education that does not require unusual institutional features. The underlying logic — using observable behavior that reveals unobserved selection tendencies — may generalize to value-added settings beyond education where agents select on unobserved characteristic that may be correlated with some observable behavior. Second, we apply our method across a broad, nationally representative sample of institutions and subjects, showing that it addresses selection across diverse university contexts. Third, we provide new evidence on the importance of instructors, including the first direct estimates of instructor value-added to earnings and new evidence on the mechanisms, particularly major choice, through which those effects operate.

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.[3] Additionally, we estimate value-added for instructors across a broader range of courses than previous studies and extend the analysis to include value-added to earnings. Our results align with the existing literature in our finding of large variation in instructor value-added to student achievement. We additionally extend analysis by Carrell and West (2010) and DeVlieger et al. (2018) by comparing instructor value-added to student evaluations of instructor quality. In contrast to the finding in Carrell and West (2010) that instructors who raise student scores in their own courses tend to receive high evaluation scores but have low value-added to student grades in subsequent courses, we find that student evaluations are positively correlated with value-added to next-semester GPA but are uncorrelated with value-added to earnings.

Contemporaneous work by Shi (2025) examines long-run impacts of high value-added instructors in developmental courses at community colleges, using the same Texas dataset. Both papers find that instructors matter for long-term outcomes and that observable instructor characteristics explain little variation.[4]

We also build on a much larger literature that measures instructor value-added to test scores in K-12 education. This literature uses value-added methods to demonstrate that teachers in primary and secondary schools have causal impacts on student outcomes across a variety of settings.[5] 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 conditional randomization

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

[4]Shi (2025) focuses on first-year students in developmental courses at community colleges, a setting chosen in part to mitigate concerns about student-instructor matching. We study a broader set of instructors at four-year institutions and address selection by conditioning on students' course histories across multiple semesters. Shi estimates instructor value-added as residual pass rates; we estimate value-added directly with respect to long-run outcomes. Differences in effect sizes likely reflect greater earnings variance among four-year graduates and lower baseline completion rates among two-year students.

[5]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).

policy to validate estimates from non-randomized data. Chetty et al. (2014a) develop forecast bias tests for selection on omitted observables in New York schools, which we apply for omitted observables in both of our settings. 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.

2.1 Data generating process

Consider a generic achievement measure A^*.[6] 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

[6]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.

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

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

where X_{it} captures student i's background characteristics, C_{jct} are characteristics of other students taking class c with instructor j,[7] 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.[8] 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 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^*.

[7] 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.

[8] 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.

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.[9] Estimating instructor value-added to lifetime earnings

[9]Students could be choosing instructors in other ways as well. For example, all students might choose

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**.[10] 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 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

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.

[10] 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.

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 B 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. Our outcome of interest at Purdue is next-semester GPA, as a frequently observed summary measure of students' academic achievement.

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.[11] This assignment mechanism generates conditional random assignment: by construction, students with the same or similar submitted preference lists have the same or similar probabilities of being assigned to any given instructor.[12] 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, were exempt from algorithmic assignment and are excluded from our analysis. Appendix C 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.

[11] 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.

[12] 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.

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 35 universities. In Texas, we also use next-semester GPA as our primary academic outcome.

These transcripts are linked to the state's unemployment insurance system, allowing us to track quarterly earnings for students who remain employed in Texas.[13] Our primary earnings measure is log quarterly earnings six-years post-matriculation.[14] 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.[15] Second, we can explore the relationship between instructor characteristics and value-added. Third, the 35 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 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.

[13] 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.

[14] 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.

[15] 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.

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.[16] 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-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 instructor, 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

[16] 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.

semester as another instructor, without requiring a normality assumption on the distributions of value-added or errors.[17]

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.[18]

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.[19]

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.[20]

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 during the same semester across instructors. 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

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

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

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

[20] 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.

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.[21]

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.1 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 Marketing, 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

[21] Appendix D.2 describes results from a robustness test where we vary the divergence threshold, which are similar to the results from our preferred threshold. Additionally, see Appendix B.4 for additional implementation details. Appendix Table A-7 summarizes the resulting group structure: the mean number of groups per course-semester is approximately 11, with an average of 5 students per group.

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.[22] 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 as *conditionally random* or *benchmark* estimates.

The conditional random variation serves two purposes: it provides a "gold standard" comparison for our course history estimates, and it supplies an additional source of identifying variation for forecast bias tests in 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

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

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 post-matriculation, 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 un-

observably 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.

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:

$$A^*_{ijct} = \delta^{ind}\widehat{\mu_{jt}} + X_{it}\beta^{ind} + C_{jct}\gamma^{ind} + \rho^{ind}_c + \lambda^{ind}_t + \psi^{ind}_{ijct}, \tag{11}$$

where X_{it} and C_{jct} include the same controls as in the initial value-added estimation, and we additionally include the course history group fixed effect H_{ict} when testing course history value-added, or 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 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 residual 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.[23] In Texas, where richer background data are available, we additionally check 9th-grade standardized test percentiles 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:

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

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

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 35 universities. Across both settings and all tests, our criterion for a credible value-added estimator is that estimates of forecast bias 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 0.282, 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 0.861, 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	0.282	0.861
	(0.148)	(0.175)
Preference Similarity Controls	X	X
Observations	132,186	132,109

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 similarly small coefficient. The coefficient on the course history estimate is slightly larger, but is not statistically distinguishable from zero. This test suggests that twice-lagged GPA is not a major source of omitted variable bias for any of our methods.

Table 3. Prediction forecast bias test at Purdue

Panel A: 2018-2023	Lagged Achievement	Course Histories	Cond. Random
Value-added Estimate	0.014	-0.044	-0.018
	(0.023)	(0.057)	(0.004)
Observations	130,404	100,428	128,362
Panel B: 2011-2023	Lagged Achievement	Course Histories	
Value-added Estimate	-0.016	0.072	
	(0.014)	(0.055)	
Observations	702,188	690,160	

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.

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 achieve-

ment value-added has coefficients far from 1 — 0.228 in the post-policy period and -0.012 in the full sample — indicating severe forecast bias. Course history value-added has coefficients near 1 (0.812 and 0.833), as does conditionally random value-added (0.790 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 predict almost none of the out-of-sample variation in student outcomes.

Table 4. Individual forecast bias test at Purdue

Panel A: 2018-2023	Lagged Achievement	Course Histories	Course Preferences
Value-added Estimate	-0.012	0.812	0.790
	(0.076)	(0.090)	(0.070)
Observations	277,996	226,290	220,087
Panel B: 2011-2023	Lagged Achievement	Course Histories	
Value-added Estimate	0.228	0.833	
	(0.084)	(0.042)	
Observations	1,132,421	1,126,843	

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.

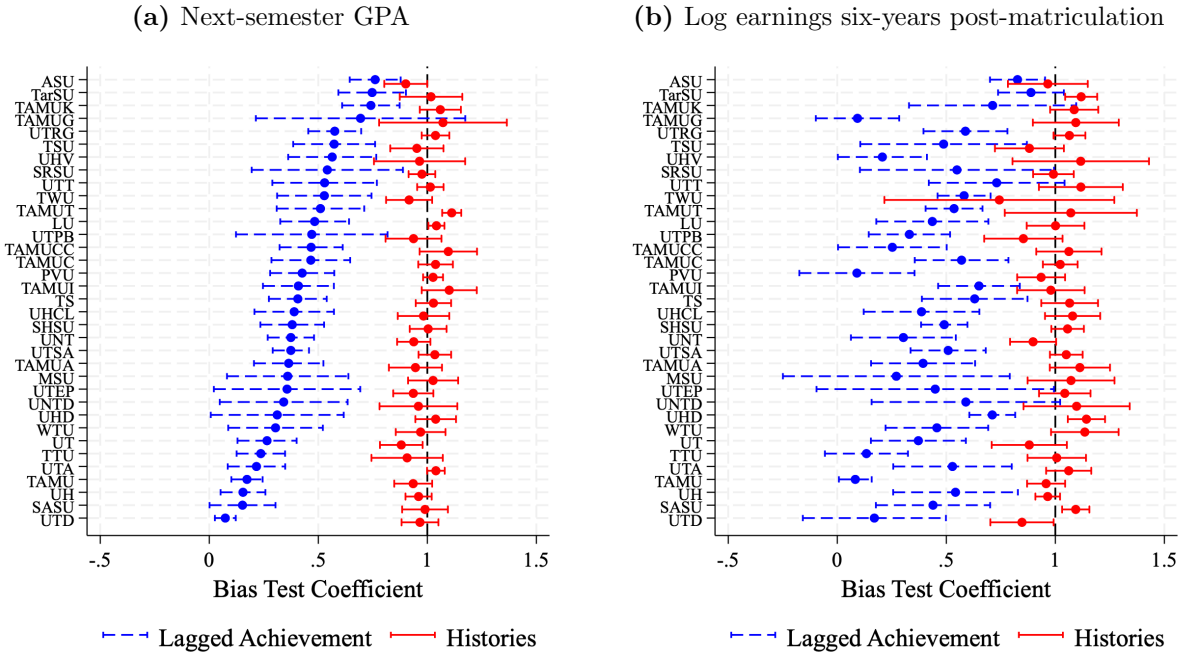
5.3 Forecast bias results at Texas universities

We now assess whether course histories address selection bias at a broader set of institutions. Unlike Purdue, the 35 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 post-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.

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.

Much like at Purdue, 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.[24] Appendix figures A-4 and A-5 present estimates for forecast bias coefficients for the universities in our Texas sample. 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 at some universities. The bias is slightly larger for value-added with lagged achievement, (average coefficient of 0.02 for lagged achievement and 0.01 for course histories), but the magnitudes are small for both methods. Value-added to earnings shows no bias from twice-lagged GPA under either method.

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

5.4 Summary

Across three forecast bias tests at Purdue and two tests at 35 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 even tests that do not share its identifying assumption, with a coefficient of 0.282 in the policy test, and coefficients similarly far from one in individual tests. Course history value-added performs far better, with coefficients of 0.861 on the policy test and coefficients comparable to the 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 35 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 average standard deviations of value-added to next-semester GPA and future earnings across subjects at all Texas universities. The standard deviations represent the average 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 0.05 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.[25] Instructors also affect future earnings: an instructor with 1 s.d. higher value-added to earnings increases students' log earnings by 0.05 log points six years post-matriculation.[26]

[25]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.

[26]This estimate includes only students with non-zero earnings in at least one quarter during the sixth

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.

Table 5. Average standard deviations of value-added distributions

Value-added Outcome	Next Semester GPA	Earnings
Mean of SD	0.0501	0.0502
Median of SD	0.0368	0.0265
Number of Programs	1,339	1,339

Notes: Standard deviations are the square root of variances estimated within subject and institution a split-sample procedure described in Appendix 4.1. Means and medians were then taken across subjects and institutions, weighted by student-course-period observations.

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.[27] The R^2 values are 0.014 for value-added to GPA and 0.005 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.[28] 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 differ-

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.

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

[28]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).

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	0.006	0.029*
	(0.015)	(0.016)
N	62,325	59,066
F-test	0.730	0.285
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)
Observations	62,388	59,122
Equality Test	0.091	0.362
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.

ences 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, contingent instructors have significantly higher value-added than assistant professors.[29] This finding is relevant for the policy analysis in Section 7, where we 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 only cleanly separates channels under strong assumptions. 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

[29] Previous work has found contingent instructors improve (Bettinger and Long, 2010; Figlio et al., 2015), reduce (Ehrenberg and Zhang, 2005), or have no effect on student achievement relative to tenure-track faculty.

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.

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.

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-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.[30] Adding instructor leniency produces a similar reduction.[31] In contrast,

[30] 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.

[31] 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.

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 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.[32] 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

Universities make consequential personnel decisions — hiring, promotion, retention, tenure — using student evaluations because they lack objective alternatives. If evaluations were

[32]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.

valid measures of teaching effectiveness, this practice would be defensible. We show in this section that they are not, insofar that they do not strongly predict instructor impacts on GPA or earnings. Student evaluations are weakly correlated with value-added to GPA, uncorrelated with value-added to earnings, and more strongly predicted by instructor grading leniency than by either value-added measure. Yet evaluations strongly predict retention decisions, particularly at the bottom of the distribution. We close the section by quantifying the earnings gains from a counterfactual retention policy motivated by work in Hanushek (2009) and Chetty et al. (2014b) that uses value-added in place of evaluations.

7.1 Correlation between value-added and student evaluations

Using data from four Texas universities, we investigate whether student evaluations align with value-added measures of instructor quality. We scraped teaching evaluations from the University of Texas at Austin, the University of Texas at Dallas, Texas Tech University, and Sam Houston State University. The surveys vary across institutions and years, so we focus on three categories of questions that appeared in all years and institutions: "Instructor Score" and "Course Score" capture students' general impressions of the instructor and the course; "Teaching Score" specifically assesses the instructor's effectiveness in conveying material. We standardize average evaluation scores within subject and institution.[33]

The first two columns of Table 8 report regressions of value-added on evaluation scores, conditional on institution-by-subject fixed effects.[34] Value-added to GPA is significantly correlated with evaluations, but magnitudes are small: instructors with 1 s.d. better Teaching Score evaluations have 0.068 s.d. higher value-added to GPA. Value-added to earnings shows essentially no correlation with any evaluation measure, with precisely estimated near-zero coefficients.

Column 3 of Table 8 examines the relationship between evaluations and instructor grading leniency, defined as the difference between an instructor's average grades and the average given by other instructors teaching the same course. Across all three evaluation categories, evaluations are roughly three times as strongly correlated with leniency as with value-added to GPA. An instructor with 1 s.d. higher overall scores has 0.243 s.d. higher leniency. This finding is consistent with prior evidence that evaluations distort grading incentives (Nelson and Lynch, 1984; Eiszler, 2002). Students appear to reward instructors who grade easily, not instructors who teach effectively.[35]

[33] For details on question categorization, see Appendix E.

[34] For comparability, values in each column are normalized within subject and institution.

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

Table 8. Comparison of instructor value-added to student evaluations

	Value-Added		
	GPA	Earnings	Leniency
	(1)	(2)	(3)
Instructor Score	0.068***	0.014	0.243***
	(0.013)	(0.018)	(0.025)
Observations	8,630	8,264	9,110
Course Score	0.063***	0.019	0.244***
	(0.015)	(0.020)	(0.022)
Observations	7,284	7,158	7,692
Teaching Score	0.086***	0.011	0.269***
	(0.016)	(0.018)	(0.028)
Observations	8,136	7,788	8,579

Notes: Estimates are from separate regressions of instructor and course quality measures, based on student evaluations, on instructor value-added or leniency, including institution-subject fixed effects. An observation is an instructor at one of the institutions where we observe student evaluations: UT Austin, UT Dallas, Texas Tech, and Sam Houston State University. 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. Standard errors are clustered at the institution-subject level.

If evaluations were merely imperfect signals that universities ignored, this disconnect would be a curiosity. The next subsection shows that evaluations actively shape personnel decisions.

7.2 Student evaluations and instructor retention

To establish that evaluations enter personnel decisions in practice, we document the relationship between evaluations and an instructor's probability of being retained the following year. We focus on contingent instructors, who are less likely than tenure-track faculty to have longer-term employment contracts that restrict termination.

For a contingent instructor j teaching in subject s at university u in year y, we define retention $R_{j,s,u,y+1}$ as an indicator for whether the instructor teaches again in the same subject at the same university the following year. We predict retention with the average

"Instructor Score," $S_{j,s,u}$,[36] both linearly and non-linearly:

$$R_{j,s,u,y+1} = S_{j,s,u} + \lambda_{su} + \gamma_y + \epsilon_{j,s,u,y} \tag{16}$$

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

where $S^v_{j,s,u}$ is an indicator for instructor j being in the vth vigintile of evaluations within subject and university (with the top vigintile as reference), and λ_{su} and γ_y are subject-by-institution and year fixed effects.

Evaluations strongly predict retention. Appendix Table A-5 reports the linear estimates:[37] A 1 s.d. increase in evaluations is associated with an 3.3 percentage points increase in the likelihood of retention.[38] Neither value-added measure significantly predicts retention; the value-added coefficients are an order of magnitude smaller than the corresponding estimates for evaluation scores.

The relationship between student evaluation scores and retention is non-linear: retention probabilities are roughly flat across most of the evaluation distribution, but drop sharply for instructors in the bottom 5%. Appendix Table A-6 shows instructors in the bottom 5% are about 14 percentage points less likely to be retained than instructors in the top 5%, and substantially less likely to be retained than the next-lowest vigintile. This discontinuity suggests that universities use evaluations as a deselection tool for the lowest-rated instructors.

Given that evaluations inform retention but are only weakly correlated with the learning and earning outcomes universities may care about, replacing evaluation-based deselection with a value-added-based alternative should improve outcomes. We quantify this in the following section.

7.3 Counterfactual deselection exercise

To evaluate the possible benefits estimates of value-added could bring to students, we consider policies a university could undertake to improve student earnings by making personnel decisions using value-added estimates instead of student evaluations.[39] We consider policies that focus on deselecting the bottom 5% of instructors in the value-added distribution.

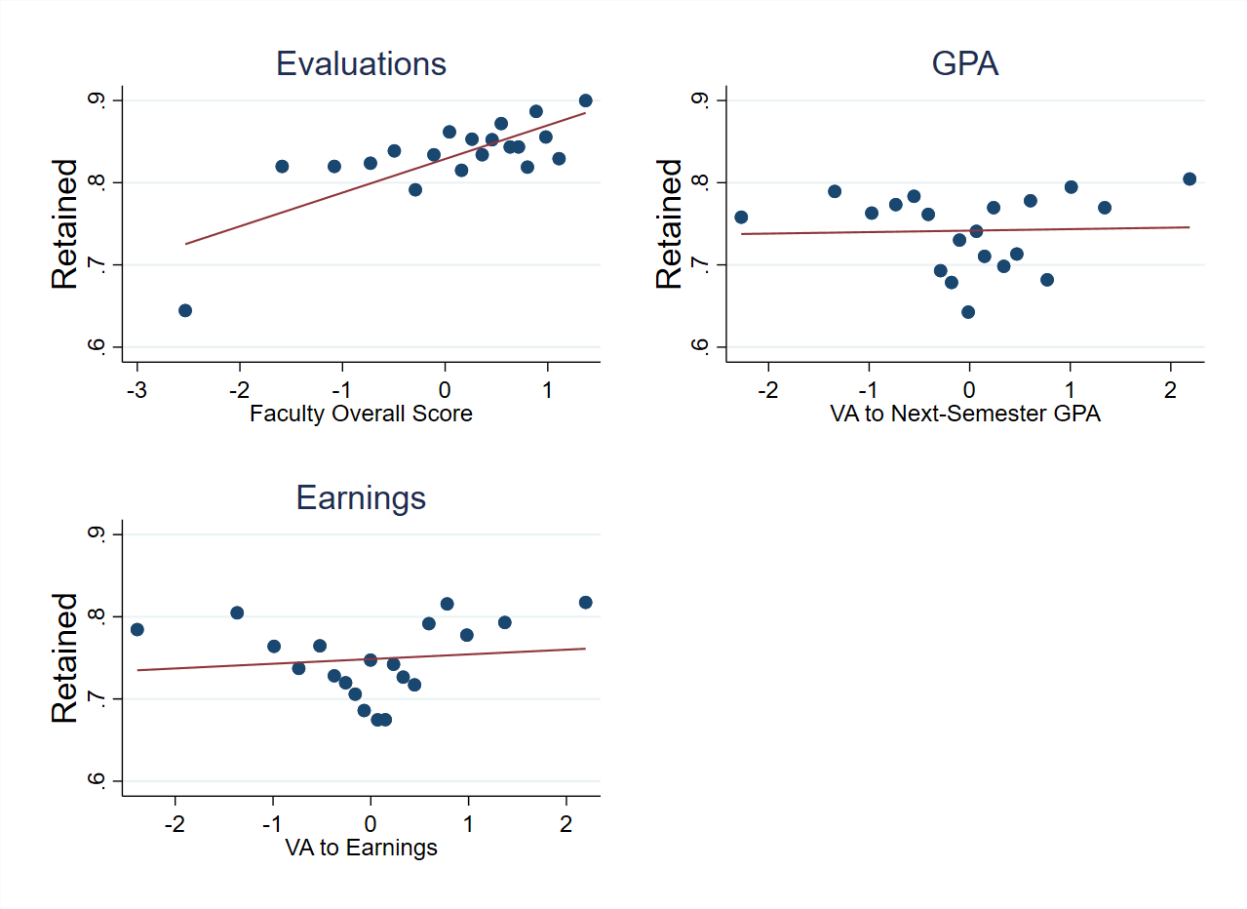
[36] We use the time-fixed average of student evaluations to be consistent with our time-fixed measures of value-added. Results using time-varying evaluation score averages are similar.

[37] Appendix Figure 2 represents the evaluation-retention relationship in a binscatter plot.

[38] Approximately 5% of the mean, or 10% of a standard deviation.

[39] We take increasing student earnings as the university's objective, returning to the choice of objective at the end of this section.

Figure 2. Retention rates and instructor quality



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.

These exercises are inspired both by the deselection proposal in Hanushek (2009) and by our empirical finding that, at four universities in Texas, instructors in the bottom 5% of the evaluation distribution do indeed seem to be retained at much lower rates than other instructors.

Though the suggestion in Hanushek (2009) is to deselect every instructor in the bottom 5% of the value-added distribution, 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 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.

We therefore make two restrictions on the counterfactual policies considered: first, we focus on deselection of contingent instructors and second, we focus on counterfactual policies which approximate the current number of annual replacements but bases retention decisions on instructors' value-added rather than student evaluations.[40]

We consider two versions of the counterfactual policy exercise that differ in their use of value-added measures. Our *infeasible counterfactual* deselects on value-added to earnings directly. This policy provides an upper bound on potential gains, but comes with important limitations. First, such a deselection tool would only be available to institutions with access to earnings data. Second, the method cannot be implemented in real time: estimating value-added to earnings requires observing students' earnings several years after they take a course, so a university would have to wait nearly a decade after an instructor begins teaching before assessing their performance.

The *feasible counterfactual* deselects on a proxy measure that universities could compute in close-to-real time: the value-added to the average earnings of a student's graduating major. The motivation for this proxy follows from Section 6.3, which showed that nearly half of the instructor effect on earnings operates through major choice. If instructors affect earnings primarily by redirecting students into high-earning fields, then a measure capturing instructor effects on major choice should approximate the true object of interest, but with a shorter lag than earnings after a student has entered the labor force.

We estimate value-added to average major earnings by calculating average earnings at the 4-digit CIP code level, then using these averages as the outcome in our standard course history value-added estimation procedure.[41]

Both counterfactuals apply the same replacement rate to bottom 5% instructors that universities currently apply when deselecting based on evaluations. We estimate this empirical deselection rate as the gap between the actual retention rate of bottom 5% instructors and the rate predicted by extrapolating the linear retention-evaluation relationship from the top 95% of instructors. Specifically, within each institution, we regress retention on normalized evaluation scores and subject fixed effects using only the top 95% of instructors, then evaluate the prediction at the average evaluation score of the bottom 5%. The difference between this prediction and the observed retention rate for instructors in the bottom 5% is

[40] Furthermore, to account for noise in the value-added estimates, we evaluate policies using shrunk empirical Bayes value-added estimates.

[41] This alternative value-added measure performs comparably to our other value-added measures on forecast bias tests. Results available from the authors upon request.

our empirical deselection rate, $\hat{r}_{su}$.[42]

For a student i in the course of a bottom 5% instructor j in subject s at university u, we calculate the gains from each counterfactual policy as

$$G_{ijsu} = \hat{r}_{su} \times Income_i \times \left(-\hat{\mu}_{js}^{eb}\right),$$

where $Income_i$ is student i's winsorized earnings six years post-matriculation and $\hat{\mu}_{js}^{eb}$ is the empirical Bayes estimate of instructor j's value-added to log earnings. For the infeasible counterfactual, the bottom 5% is selected using value-added to earnings, and for the feasible counterfactual, the bottom 5% is selected using value-added to average major earnings. We sum gains across all replacements for each student and calculate the average gain for affected students.

Both the infeasible and feasible counterfactual deselection policies produce meaningful gains to student earnings compared to an evaluation-based status quo. The infeasible policy, deselecting on value-added to earnings, would increase quarterly earnings of affected students by about \$395, which is 3.9% of average quarterly earnings for this group. The feasible policy, deselecting on value-added to average major earnings, generates gains approximately one-third of the infeasible policy: \$132, representing 1.3% of average quarterly earnings for this group of students.

7.4 Discussion

The analysis produces two insights. First, the gap between the infeasible and feasible policies tells us the cost of using a proxy: institutions lose roughly two-thirds of the potential gains from the ideal policy by switching to a measure that can be computed more quickly (1.3% vs 3.9%). Second, even the feasible policy generates economically meaningful gains, comparable in magnitude to the lifetime earnings effects in Chetty et al. (2014b). These results suggest that universities have substantial room to improve student outcomes by reorienting personnel decisions away from evaluations and toward outcome-based measures of instructor quality — even when the ideal outcome measure is unavailable in real time.

Our analysis takes student earnings as the university's objective, following similar exercises in the K-12 literature. This choice is not innocuous. Universities — particularly selective research institutions and liberal arts colleges — plausibly value outcomes beyond earnings, including intellectual development, civic engagement, and matching students to fields of study in which they have comparative advantage.

[42] In both counterfactuals, we restrict to subjects with nonzero estimated variance of value-added to earnings. When selecting on value-added to major earnings, we also restrict to subjects with nonzero estimated variance of value-added to major earnings.

The feasible counterfactual brings this tension into sharp relief. Deselecting instructors based on value-added to average major earnings rewards two types of instructors: those in high-earning fields who retain students within their field, and those in low-earning fields who redirect students out. To the extent that universities value breadth of instruction across disciplines, or to the extent that students are best served by sorting into fields in which they have genuine interest and aptitude, the feasible policy creates a conflict between producing high-earning graduates and supporting the full range of academic missions that universities have traditionally pursued.

Alternative outcomes — such as value-added to on-time graduation, degree completion, or field-specific earnings — could be used instead. Each carries its own distortions to instructor and institutional incentives, and variation in value-added may be smaller across instructors for outcomes that are less dispersed than earnings. Our framework is agnostic about the choice of outcome; the course history method can be applied to any measure a university wishes to target. The question of which outcomes universities *should* target is beyond the scope of this paper, but one that merits serious attention as institutions consider adopting value-added measures in personnel decisions.

8 Conclusion

Postsecondary instructors shape students' academic trajectories and long-run labor market outcomes. Yet, until now, credible estimates of their effectiveness have been confined to a handful of institutions with unusual enrollment policies, and measured only for narrow sets of courses. Most universities, lacking objective alternatives, rely on student evaluations to make consequential personnel decisions. We show that this practice has real costs: evaluations are more strongly predicted by grading leniency than by teaching effectiveness, and are essentially uncorrelated with the outcome — earnings — that universities may most want to improve.

In this paper, we propose and validate a method that addresses this gap. Students' course histories reveal otherwise unobservable information about student types that, when controlled for, substantially reduces forecast bias in value-added estimates, performing nearly as well as estimates that exploit actual randomization of students to instructors. Applying this method across 35 Texas public universities, we find that individual instructors matter substantially: moving to an instructor one standard deviation above the mean within a course increases next-semester GPA by 0.05 points and earnings six years post-matriculation by 0.05 log points. Observable instructor characteristics explain almost none of this variation, meaning value-added is genuinely new information that institutions cannot recover from

CVs or salary schedules. A counterfactual retention policy using value-added rather than evaluations to deselect the bottom 5% of contingent instructors could increase quarterly earnings of affected students by nearly 4%.

Our two main outcomes — next-semester GPA and earnings — are intentionally complementary. GPA is available in real time and directly relevant to graduation, major eligibility, and graduate school prospects, but it is internal to the university and shaped by grading norms that vary across subjects and instructors. Earnings are market-determined, broadly comparable, and directly tied to student welfare, but realized with a lag that limits their usefulness for contemporaneous personnel decisions. Decomposition analysis suggests the two outcomes operate through related but distinct channels: roughly half of the instructor effect on GPA flows through students' subsequent course and instructor choices, while roughly half of the instructor effect on earnings flows through graduating major — a finding consistent with the large literature documenting substantial returns to field of study.

These findings point to a rich agenda for future work. The course history method can be applied to intermediate outcomes realized between GPA and earnings — graduation rates, time to degree, major attainment, or field-specific earnings — each of which could give institutions a more actionable and timely signal of instructor quality while providing researchers deeper insight into the mechanisms we document here.[43] The method also opens the door to studying heterogeneous instructor effects — by student race, gender, or ability — a question of both scientific and policy importance that the scale of our data is well-suited to address.[44] More broadly, the underlying logic of using observable behavior to proxy for unobserved selection tendencies may travel to other value-added settings — in healthcare, management, or financial advising — where agents and clients match on characteristics that are unobserved by the econometrician but revealed through behavior.

Universities have long lacked the tools to identify their most effective instructors. This paper provides a method, validation, and a set of results that together suggest the returns to getting this right are substantial.

[43] In K-12, Jackson (2018) shows that the same strategies used to estimate value-added to test scores can produce unbiased estimates of value-added to non-cognitive outcomes, suggesting that the approach extends naturally beyond traditional achievement measures.

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

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. Texas 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 (1)	VA to Earnings (2)
Full Professor	-0.182	0.482***
	(0.128)	(0.131)
Associate Professor	-0.167	0.481***
	(0.134)	(0.132)
Assistant Professor	-0.169	0.449**
	(0.129)	(0.133)
Contingent	-0.158	0.462***
	(0.118)	(0.124)
Asian	-0.023	0.017
	(0.015)	(0.014)
Black	-0.030**	-0.021
	(0.014)	(0.018)
Hispanic	0.013	-0.004
	(0.014)	(0.014)
Female	-0.003	-0.006
	(0.009)	(0.008)
International	-0.052**	0.000
	(0.020)	(0.018)
Log Average Salary	0.028**	-0.039**
	(0.011)	(0.011)
Age	0.008	0.010
	(0.005)	(0.006)
Fraction Upper Level	-0.062**	0.024
	(0.020)	(0.018)
Observations	61,854	58,690
R^2	0.014	0.005

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. Linear retention of contingent instructors

	Retained (1)	Retained (2)
Instructor Score	0.033***	0.025*
	(0.007)	(0.014)
Value-Added Controls		X
Observations	13,994	5,517

Notes: Estimates are from regressions of a retention indicator on normalized average student evaluations with university by subject and year fixed effects, using instructors from UT Austin, UT Dallas, Texas Tech, and Sam Houston State. Observations are at the instructor-year level, restricted to contingent instructors. Column 2 controls for normalized value-added to next-semester GPA and earnings and restricts to instructors with estimates for both value-added to GPA and value-added to earnings. Standard errors are clustered at the university-subject level.

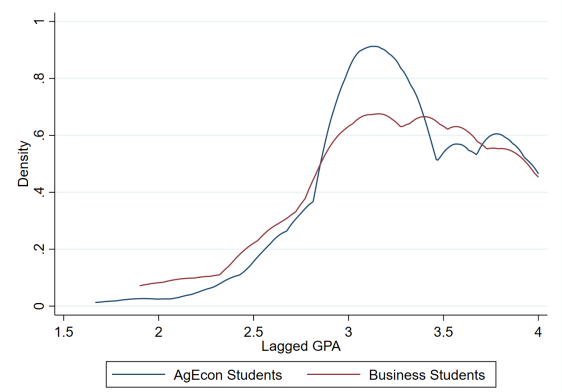
Table A-6. Non-linear retention of contingent instructors

	Retained (1)	Retained (2)
Instructor Score Vigintile 1	-0.145***	-0.083*
	(0.035)	(0.047)
Instructor Score Vigintile 2	-0.041	-0.057
	(0.031)	(0.048)
Instructor Score Vigintile 3	-0.021	-0.020
	(0.028)	(0.055)
Instructor Score Vigintile 4	0.028	0.048
	(0.027)	(0.042)
Instructor Score Vigintile 5	-0.013	0.011
	(0.031)	(0.040)
Instructor Score Vigintile 6	-0.030	-0.035
	(0.030)	(0.049)
Instructor Score Vigintile 7	-0.008	-0.017
	(0.028)	(0.044)
Instructor Score Vigintile 8	-0.003	-0.053
	(0.029)	(0.046)
Instructor Score Vigintile 9	0.023	0.024
	(0.027)	(0.045)
Instructor Score Vigintile 10	0.001	0.008
	(0.031)	(0.041)
Instructor Score Vigintile 11	0.016	-0.011
	(0.030)	(0.054)
Instructor Score Vigintile 12	-0.011	-0.053
	(0.023)	(0.043)
Instructor Score Vigintile 13	-0.039	-0.085
	(0.048)	(0.074)
Instructor Score Vigintile 14	-0.002	-0.044
	(0.023)	(0.040)
Instructor Score Vigintile 15	0.013	-0.012
	(0.025)	(0.037)
Instructor Score Vigintile 16	0.029	0.005
	(0.022)	(0.025)
Instructor Score Vigintile 17	0.010	0.015
	(0.024)	(0.035)
Instructor Score Vigintile 18	0.029	-0.006
	(0.020)	(0.028)
Instructor Score Vigintile 19	0.031	0.036
	(0.025)	(0.035)
Value-Added Controls		X
Observations	13,979	4,786

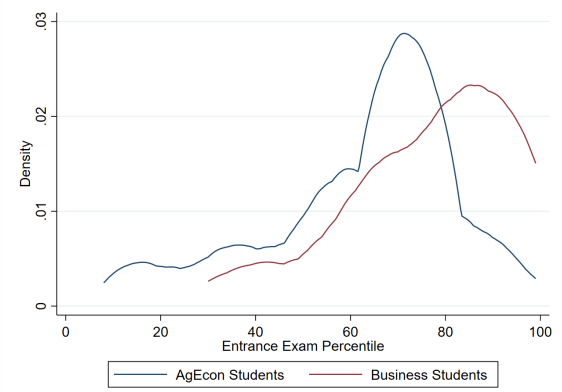
Notes: Estimates are from regressions of a retention indicator on vigintiles of average student evaluations with university by subject and year fixed effects. Observations are at the instructor-year level, restricted to contingent instructors. Column 2 controls for vigintiles of value-added to next-semester GPA and earnings and restricts to instructors with estimates for both value-added to GPA and value-added to earnings. Standard errors are clustered at the university-subject level.

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

(a) Lagged GPA

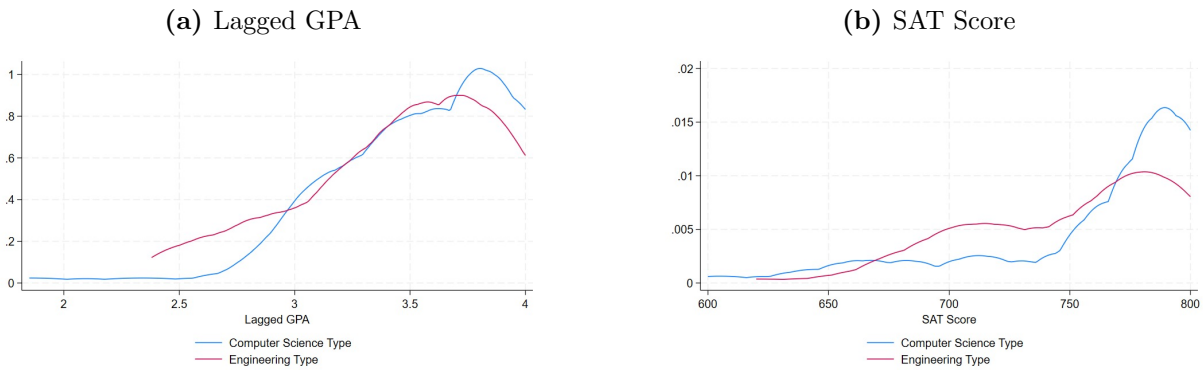


(b) Entrance Exam Percentile



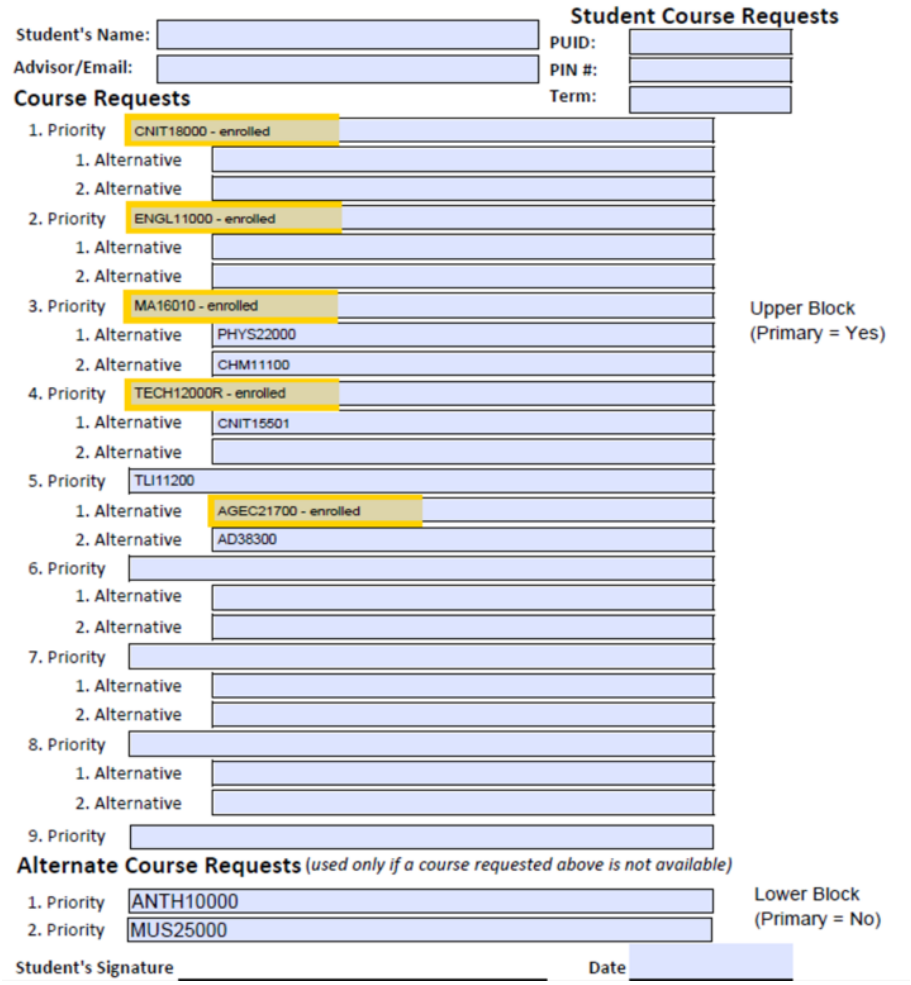
Notes: This figure shows smoothed kernel density plots for students in two different course history groups taking Intermediate Micro in the same semester at Texas A&M.

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



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)

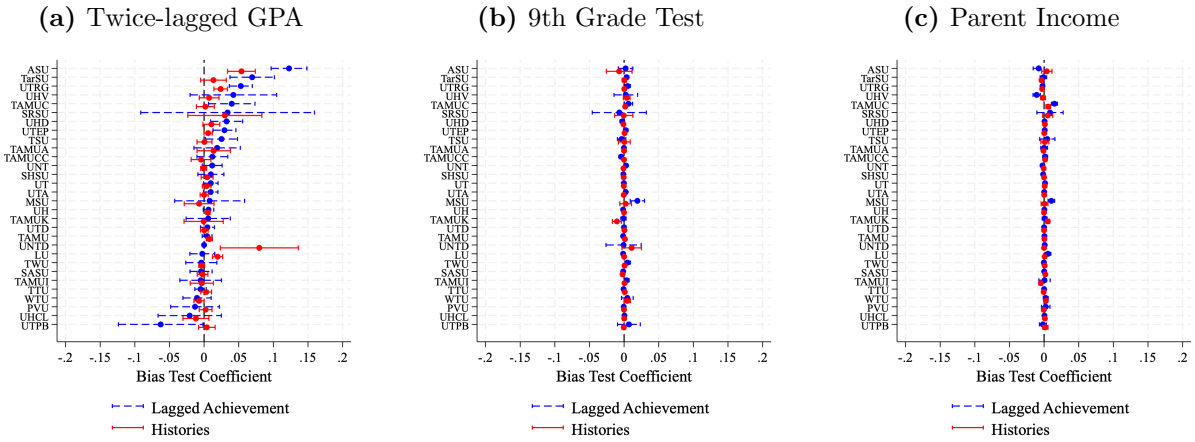
Alternate Course Requests *(used only if a course requested above is not available)*

1. Priority: ANTH10000
2. Priority: MUS25000

Lower Block (Primary = No)

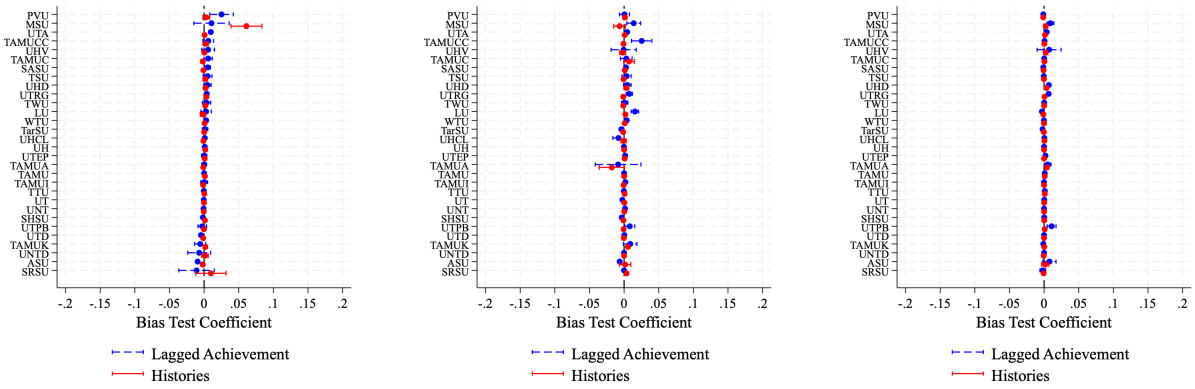
Student's Signature _____ Date _____

Figure A-4. Prediction forecast bias test of value-added to GPA at Texas universities



Notes: Each point is a point estimate in a regression of predicted residual GPA on value-added, with 95% confidence intervals. Subfigure (a) predicts residual GPA using twice-lagged GPA, subfigure (b) predicts using 9th-grade test percentile, and subfigure (c) predicts using parent income. Value-added is estimated using the subsample of students missing nonmissing values of each predictor. Standard errors clustered at the subject level.

Figure A-5. Prediction forecast bias test of value-added to earnings at Texas universities



Notes: Each point is a point estimate in a regression of predicted residual log earnings six years post entry on value-added, with 95% confidence intervals. Subfigure (a) predicts residual log earnings using twice-lagged GPA, subfigure (b) predicts using 9th-grade test percentile, and subfigure (c) predicts using parent income. Value-added is estimated using the subsample of students missing nonmissing values of each predictor. Standard errors clustered at the subject level.

B Estimation Details

B.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

$$A^*_{ijct} = X_{it}\beta^s + C_{jct}\gamma + \rho^s_c + \lambda^s_t + \mu^s_j + \epsilon_{ijct}, \tag{18}$$

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.

We then estimate the variance of the error term using a simple method-of-moments estimator following Gilraine et al. (2020):

$$\hat{\sigma}^2 = \frac{\sum_j \sum_t \sum_i (A_{ijct} - A_{jt})}{\sum_j \sum_t (N_{jt} - 1)},$$

where A_{ijct} are the residuals from Equation 7, 8, or 9, and A_{jt} is the average of the residuals for instructor j in period t, and N_{jt} is the number of student-course-semester observations taught by instructor j in period t.

B.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 Equations 7, 8, or 9, and obtain residuals A_{ijct}. We then estimate σ^2_μ and σ^2_ϵ as discussed in Appendix B.1. With these estimates in hand, we construct average residuals $\overline{A}_{jt} = \frac{1}{N_{jt}} \sum_c \sum_i A_{ijct}$, where N_{jt} is the number of students taught by instructor j in period t in a given subject.

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

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

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:

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

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 $\hat{\sigma}^2_\mu$ is smaller relative to $\hat{\sigma}^2_\epsilon$, since in these subjects, measurement error is more of a problem.

B.3 Jackknife empirical Bayes estimates

When regressing an outcome on estimates of value-added, such as in the forecast bias test in Equation 11, 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.

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. This results in time-varying weighted sums m_{jt}, which we multiply by the updated shrinkage term, using the new weights h_{jt} that leave out t. Finally, we obtain jackknife empirical Bayes estimates $\hat{\mu}_{jt}$.

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

B.4 Hierarchical clustering

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 $\mathbf{H_{ict}}$[45], which we define as the vector of indicators for each course offered at an institution, where an entry is 1 if a student has taken the course corresponding to the entry during a period $t' \leq t$ and a zero otherwise. The methods described are also used to group students at Purdue based on their top six course preferences, similarly encoded in the vector P_{ict}, but without loss of generality, we will focus on histories.

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

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

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

[45] 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

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.

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

Table A-7. Course history similarity group statistics

	Mean (1)	Median (2)
Groups per Course	10.892	4
Students per Group	5.193	2
Fraction Ungrouped	0.299	.
Fraction in Groups of ≥ 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.

C 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:

$$weight(a \in dom(R)) = 0.9^{prior(R)} \times 0.5^{alt(a)}, \tag{19}$$

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.

D Robustness Tests

D.1 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 8. Forecast bias test results and variance estimates of value-added estimates

including ungrouped students are very similar to results for our preferred estimates. Results are available from the authors upon request.

D.2 Alternative history clusters

We also test robustness of choice of divergence level by defining course history groups for a variety of divergence levels and re-estimating value-added. Most choices of divergence level perform well on forecast bias tests, but our preferred choice of mean divergence outperforms other choices on forecast bias tests. Estimates using each different choice of divergence also gives similar variances to our choice of mean divergence. Results are available from the authors upon request.

E Student Evaluation Questions

Table A-8 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.

Table A-8. Evaluation Question Categorization

Institution	Evaluation prompt	Years available
Instructor score		
Sam Houston	Overall, I rate this instructor as excellent	2005–2023
Texas Tech	Overall this instructor was effective	2009–2014
Texas Tech	Overall, the instructor was an effective teacher	2014–2023
UT Austin	Instructor quality (1–5)	2005–2022
Course score		
Sam Houston	Overall, I rate this course as excellent	2005–2023
Texas Tech	Overall this course was a valuable learning experience	2009–2023
UT Austin	Course quality (1–5)	2005–2022
UT Tyler	The instructor clearly defined and explained the course objectives and expectations	2018–2023
UT Tyler	The instructor was prepared for each instructional activity	2018–2023
UT Tyler	The instructor communicated information effectively	2018–2023
Teaching score		
Sam Houston	Asked students to share ideas and experiences with others whose backgrounds and viewpoints differ from their own	2005–2023
Sam Houston	Created opportunities for students to apply course content outside the classroom	2016–2023
Sam Houston	Demonstrated the importance and significance of the subject matter	2005–2023
Sam Houston	Encouraged students to use multiple resources (e.g. Internet, library holdings, outside experts) to improve understanding	2005–2023
Sam Houston	Explained course material clearly and concisely	2005–2023
Sam Houston	Formed teams or groups to facilitate learning	2005–2023
Sam Houston	Found ways to help students answer their own questions	2005–2023
Sam Houston	Gaining a basic understanding of the subject (e.g. factual knowledge, methods, principles, generalizations, theories)	2016–2023
Sam Houston	Gave projects, tests, or assignments that required original or creative thinking	2005–2023
Sam Houston	Introduced stimulating ideas about the subject	2005–2023
Sam Houston	Involved students in hands-on projects such as research, case studies, or real-life activities	2005–2023
Sam Houston	Made it clear how each topic fit into the course	2005–2023
Sam Houston	Provided meaningful feedback on students' academic performance	2005–2023
Sam Houston	Stimulated students to intellectual effort beyond that required by most courses	2005–2023
Texas Tech	The course objectives were specified and followed by the instructor	2014–2023
Texas Tech	The instructor demonstrated knowledge of the subject	2009–2014
Texas Tech	The instructor stimulated student learning	2009–2014
Texas Tech	The instructor presented the information clearly	2009–2014
Texas Tech	The instructor emphasized the major points and concepts	2009–2014
UT Austin	The course was well organized	2005–2021
UT Austin	At this point in time I feel that this course will be (or has already been) of value to me	2005–2014
UT Austin	The course materials (e.g. text and supplemental materials) were helpful to me	2014–2021
UT Austin	Overall I learned a great deal in this course	2014–2021

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