The Economic and Social Value of a High-school Teacher: An Extended Abstract*

Jesse Bruhn† Carlos F. Avenancio-León‡ Menaka Hampole§ Rachel Moore¶

March 2026

Abstract

We leverage quasi-random variation embedded in the software algorithm that assigns students to classrooms in Texas high schools to study the causal effect of high school teachers on test scores, attendance, discipline, graduation, post-secondary attainment, civic engagement, and earnings. We document three main patterns. First, returns to experience are strongly increasing for contemporaneous test scores: relative to first-year teachers, assignment to a teacher with 16 or more years of experience raises test scores by 0.046 standard deviations. Corresponding effects on age-25 wages are also large (≈2.5p.p.) but imprecisely estimated. Second, we find evidence of consequential match effects, but they do not map cleanly across horizons: female students assigned to female teachers score about 0.10 standard deviations higher, yet analogous long-run effects on outcomes, such as 4-year degree completion, are sometimes opposite in sign. Third, classroom effects are widely dispersed, especially for earnings. The signal standard deviation of classroom effects is 0.12 student standard deviations for test scores and 0.082 log points for wages at age 25, implying that a 1-SD improvement generates roughly $2,400 in annual earnings at the sample mean. We also find that classrooms that raise test scores tend to raise adult wages. Even under conservative assumptions, a one-standard-deviation improvement in test score-based measures of classroom quality generates about $3.52 million in present discounted lifetime earnings for a single teacher-year of students. Taken together, the results highlight the social and economic importance of high-school teachers and suggest that test-score gains are informative but not exhaustive about teachers' long-run contributions to adult well-being.

*The conclusions of this research do not necessarily reflect the opinions or official position of the Texas Education Research Center, the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, and the State of Texas.

†Brown University. Corresponding Author. Email: jesse_bruhn@brown.edu.

‡University of California, San Diego.

§Yale University

¶Princeton University

1 Introduction

Every day, nearly 16 million students attend high school in the United States (National Center for Education Statistics, 2024). Much of that time is spent in the classrooms of the roughly 1.1 million teachers employed in the secondary-school labor force (U.S. Bureau of Labor Statistics, 2023). These educators are asked to do far more than raise contemporaneous test scores. High school teachers are charged with helping our children realize their full potential, both in terms of the academic skills, talents, and capacities that will benefit them in post-secondary education or the broader labor market; and also in terms of broader notions of well being, like engaging with the civic and social institutions that are key to a flourishing society. Thus, the implicit hope is that the investments these educators make in our children will not only move test scores, but also generate return on investment for outcomes like high school graduation, post-secondary attainment, civic engagement, and earnings.

Do high school teachers shape long-run social and economic well-being? Causal evidence on this question is thin. There have been exactly two studies that leveraged randomized or quasi-experimental variation to measure the subsequent impact of classroom assignment on outcomes like earnings in the US. The first, Tennessee's project STAR, was originally designed to learn the causal effect of class-size on test scores (Krueger, 1999; Krueger and Whitmore, 2001). For this study, the researchers randomized roughly 11,000 students across 1,300 classrooms in grades K-3, finding a standard deviation of classroom effects on earnings of roughly 10% (Chetty et al., 2011). The second study, Chetty et al. (2014b), leverages event-study style variation emerging from teacher's switching schools to study the link between test-score value-added and earnings for teachers in grades K-8. Other notable studies either do not observe earnings (e.g. Project MET), or otherwise rely on variation that is not design-based (e.g. value-added methods) to identify the causal impact of teachers.[1] More broadly, the literature on high-school teachers is comparatively thin and, to our knowledge, none have successfully linked to earnings.[2] This lack of evidence on highschool teachers is almost surely because teacher assignments become more complex at the secondary level. Multi-teacher configurations and the ability for students to submit preferences over expanded course offerings creates scope for complicated forms of selection that are unlikely to be accounted for by simply conditioning on a lagged test score (Jackson, 2014b; Goldhaber et al., 2013; Tartova, 2023).

We study the causal effect of classroom and teacher assignment on standardized test scores, school discipline, high school graduation, post-secondary attainment, civic engagement, and earnings. For identification, we leverage large-scale quasi-random variation embedded in the algorithm that assigns students to classrooms in Texas high schools. Classroom assignment is a complex optimization

[1] A notable example of a randomized design is Kane and Staiger (2008); however, this study has not been linked to earnings. For a broad overview of the value-added literature, see Koedel and Rockoff (2015). For more recent methodological advances, see Gilraine et al. (2020); Gilraine and Pope (2021); Mulhern and Opper (2021) and Rose et al. (2025).

[2] For papers that study high school teachers, see Aaronson et al. (2007); Goldhaber et al. (2013); Jackson (2012, 2014a, 2018); Mansfield (2015); Cook and Mansfield (2016); Liu and Loeb (2021); Backes et al. (2023) and Shah (2025). Of particular note is Lavy and Megalokonomou (2024), who use features of the scheduling system in Greece to estimate value-add on college going.

problem requiring specialized software. Within a school, thousands of students submit preferences over courses (e.g. Algebra 1, English II) and the algorithm uses those preferences to build hundreds of classrooms and assign them to teachers—all subject to section capacity limits, scheduling constraints, teacher availability, and course prerequisites. While the exact algorithms are proprietary IP of software vendors (e.g. Powerschool), preferences for teachers and peers are *not* included as fields in the user interface. Moreover, many school timetabling algorithms developed for this purpose in the operations research literature feature random student ordering or other forms of stochastic assignment (see Pillay, 2014, for an overview). Thus when two students who attend the same school are assigned the same course in the same period in the same year, the algorithm places them into parallel classrooms in a manner that is orthogonal to both the observed and unobserved characteristics of students.

To implement this strategy, we pair data from the Texas Education Research Center (ERC) with detailed records on course offerings obtained by submitting hundreds of FOIA requests to individual school districts. These course level records are necessary to recover the school-by-course-by-period-by-year cells that isolate the variation we leverage for identification. As of this writing, we have obtained complete records and merged them into the Texas ERC data for ten districts, and we are actively merging another 21 districts into the ERC records. However, delays in the disclosure process leave us with disclosed results for a single, large urban district that covers 64,355 students and 1,077 teachers across 2,708 classroom assignment strata. We expect our sample sizes to increase dramatically by the time of the Summer Institute in July, and hence the single district results we share today should be viewed as a "proof-of-concept." In addition to the standard ERC records that describe long-run outcomes such as graduation, college attendance, and earnings, on March 20, 2026, we also completed a novel merge between National Voting records (from the firm L2) and the ERC records, which will allow us to capture broader notions of adult well-being (like civic engagement) that go beyond the purely economic domain.

In our preliminary results, we document three substantive patterns. First, we find that the returns to experience are strongly increasing for contemporaneous test-score based achievement. Relative to first-year teachers, assignment to a teacher with 16 or more years of experience raises test scores by 0.046 standard deviations. By contrast, the corresponding effects on earnings at age 25 are also large ($\approx$2.5 percentage points relative to a base of $\approx$\$28,000); however, they are statistically imprecise. However, we view this as part proof-of-concept. The important point is not that experience matters only for test scores, but that the design is already powerful enough to recover a clear experience gradient in achievement while more precise evidence on the long-run margins—like earnings—await the expanded data we are currently assembling.

Second, we find evidence of student-teacher match effects; however, they do not map cleanly from short-run achievement into long-run outcomes. Female students assigned to female teachers score roughly 0.10 standard deviations higher than comparable female students assigned to male teachers, an effect that is larger than the gap between first-year and veteran teachers in our experience estimates. Yet the corresponding long-run estimates are more nuanced. For example, we find

that female teachers improve the likelihood a male student obtains a four-year college degree by 1.3p.p. (SE = 0.005), but with no evidence of a corresponding impact on female students. These findings suggest that the relevant dimensions of complementarity are richer than a simple same-sex role-model story. More generally, the research design is sufficiently well-powered, even on this limited sample, to detect match effects on long-run outcomes like degree attainment.

Our third and final finding is that the reduced-form effects of classroom assignment are widely dispersed, especially for adult earnings. To begin, we note that there is statistically precise evidence of heterogeneity: we can statistically reject that the classroom assignment effects are identically zero. Moreover, after correcting for estimation noise, the signal standard deviation of classroom effects is 0.12 student standard deviations for test scores and 0.082 log points for wages at age 25, indicating substantial dispersion. For example, at the sample mean, the latter corresponds to a roughly $2,400 causal increase in annual earnings at age 25. Moreover, we find that classrooms that raise test scores also tend to raise earnings later in life, which suggests that it is possible to use causal impacts on test scores as a short-run proxy capable of detecting classrooms that will improve long-run earnings.

We conclude with a thought experiment: what is the economic value of 1-SD improvement in test-score based measures of high school classroom quality? Under conservative assumptions about class loads, discounting, and the age profile of earnings, we find that a one-standard-deviation improvement in classroom quality implies $3.52 million in present discounted lifetime earnings for a single teacher-year of students. These estimates underscore a basic economic point: because high-school teachers teach many students each year, even moderate improvements in classroom quality can generate very large economic and social returns.

2 Data

This paper combines longitudinal administrative records from the Texas Education Research Center (ERC) with district-level course schedule files obtained under the Texas Public Information Act. The Texas ERC is a secure state longitudinal data system that links PK–12 records from the Texas Education Agency (TEA) to postsecondary records from the Texas Higher Education Coordinating Board (THECB), and quarterly wage records from the Texas Workforce Commission (TWC) allowing researchers to follow Texas students from public school into college and the labor market (Texas Education Research Center, 2026a,b; Texas Schools Project, 2026). The TEA component includes PEIMS enrollment, attendance, course completion, graduation and leaver records, assessment data, staff employment files, and teacher assignment information (Texas Education Research Center, 2026c; Texas Schools Project, 2026). We use the ERC backbone to observe student demographics, prior achievement, teacher characteristics, postsecondary enrollment and attainment, and later labor-market outcomes for cohorts observed in Texas public high schools during 2012–2019.

Implementing our research design requires an additional ingredient: exact local schedule data identifying the specific course, section, period, and teacher to which each student is assigned. Those

files are essential because the quasi-random variation exploited in the paper arises at the final step of scheduling: the assignment of students across parallel classrooms offering the same course in the same school, period, and year. The schedule files needed to recover the precise course (e.g. Honors Algebra II or English I) attached to each classroom were not available in usable form in the standard ERC extract we received for this project, so we assembled them directly from school districts via public-information requests. In most districts these files include the unique classroom identifier, general course subject (ex: Algebra I), course name (ex: Honors Algebra I), school year, and campus. We merge the district schedule files to ERC student and teacher records using the class identifier used by both the state and district administrators.

The resulting analytic file is organized at the student-course-section-year level. A student can therefore contribute multiple classroom assignments within a year, one for each eligible course. In the main analysis, we retain regularly scheduled high-school classroom observations for 8 core classes (English 1-4, Algebra I, Algebra II, Geometry, and US History) for which we can identify the school, course, period, year, and teacher, and we restrict attention to cases with at least two parallel classrooms taught by distinct teachers offering the same course in the same school, period, and year. This gives us over 140,000 student $\times$ course observations in 2,708 class periods covering 1,077 teachers and 64,355 students.

While results for only one district were disclosed in time for this submission, we have now received course data from 31 districts. Of these 31 districts, we have currently imported and merged course files for 10 districts. To give a sense of scale, we note that the 10-district merged ERC course-schedule database now covers approximately 1 million student $\times$ course observations that have at least 2 parallel classrooms—in other words, this is the current sample size we are using for estimation internally on the ERC servers. However, as noted earlier, disclosure review was not complete for the full sample at the time of this submission. Thus, the present draft reports estimates for Cypress-Fairbanks ISD (CyFair), where we had initially executed a proof of concept. For that reason, the sample we use in this extended abstract contains over 140,000 student $\times$ course observations having at least 2 parallel classrooms. Subsequent drafts will incorporate *at least* the remaining nine districts as well as potentially 21 more provided we can successful merge the newest course data into the ERC backbone. Moreover we are also actively submitting additional public-information requests to other districts to expand the sample further. Because all districts merge into the same ERC backbone, each additional course file scales the design transparently: once local schedule data are harmonized, they can be appended to the same student, teacher, postsecondary, and wage records already available through the ERC.

Our baseline covariates come from the ERC's K–12 and educator records. On the student side, we observe demographics, economic disadvantage, English-learner status, prior achievement, grade progression, and other standard administrative fields. On the teacher side, we observe experience, highest degree, gender, and race/ethnicity from PEIMS and linked educator files; when we use a continuous measure of teacher education, we map degree categories into years-of-education equivalents. The outcome measures combine short-run achievement with long-run educational and

labor-market outcomes. Test-score outcomes are constructed from statewide assessment records and standardized within exam × test-date. Postsecondary outcomes come from Texas public college records kept by the THECB (Texas Education Research Center, 2026b; Texas Schools Project, 2026). Labor-market outcomes come from quarterly unemployment-insurance wage records provided through the TWC (Texas Education Research Center, 2026a; Texas Schools Project, 2026). For our preliminary results, we study test-scores, bachelor's attainment, and earnings.

Table 1 provides summary statistics. Because disclosure of means for the full multi-district sample was not complete in time for this draft, the table compares statewide reference statistics to the disclosed CyFair sample. The statewide and district reference statistics in Panels A and C are drawn from the Texas Public Schools Explorer, which in turn reports data released through the Texas Education Agency's annual Texas Academic Performance Report (The Texas Tribune, 2026a,c,b). CyFair is a very large district, with 118,187 students in 2023–2024. Relative to the state as a whole, CyFair enrolls larger shares of Black and Asian students and a smaller share of Hispanic students, while economic disadvantage and limited-English-proficiency rates are close to statewide benchmarks. The disclosed outcome means in Panel B indicate that CyFair students in our current sample have comparatively strong academic and postsecondary outcomes. Panel C shows that CyFair teachers are slightly less experienced than the statewide average but somewhat more likely to hold a master's degree. These patterns suggest that CyFair is large and diverse but somewhat positively selected on achievement and postsecondary outcomes, a useful benchmark for interpreting the current single-district results and the likely external-validity gains from the forthcoming multi-district expansion.

[Table 1 about here.]

3 Research Design

3.1 The Teacher-to-Student Assignment Process

Building schedules for students in large high-schools is a complex optimatization problem. Large high schools ask students to submit next-year course requests in late winter or spring, typically through an online portal. District guidance in CFISD, CCISD, and Katy ISD is explicit that students choose courses—and usually alternate courses—well before the school year begins, and that these requests are used to construct the following year's master schedule (Spillane Middle School, 2026; Cy-Fair High School, 2026; Clear Creek Independent School District, 2026; Jordan High School, 2026).

Importantly, students submit preferences for specific courses, but not for specific teachers or peers. Thus, students are only able to choose whether to take Algebra I or Geometry, on-level or advanced coursework, band or athletics, dual-credit options, and endorsement-specific electives. However, converting these preferences into classroom assignments is a process that is governed by a complex set of constraints which include graduation requirements, course prerequisites, classroom

capacity limits, teacher assignments, room availability, and students other class selections (Texas Education Agency, 2026; Clear Creek Independent School District, 2026). Thus, the master-schedule problem is not a simple classroom-roster exercise; it is the problem of mapping thousands of course requests into a feasible schoolwide timetable and hence requires specialized software.

Public documentation from commercial student-information systems illustrates how these schedules are actually constructed. For example, PowerSchool's PowerScheduler documentation describes the problem as building a master schedule from student course requests, teacher schedules, teacher course requests, teacher assignments, classrooms, periods, and other constraints, and states that the system chooses among "hundreds of thousands of possibilities" to produce the schedule that satisfies the most requirements and minimizes conflicts (PowerSchool, 2026a). This description matches the broader operations-research literature, which treats school timetabling as a high-dimensional combinatorial optimization problem with many hard and soft constraints (Pillay, 2014; Tan et al., 2021). The relevant implication for our purposes is that the software incorporates courses, sections, periods, rooms, and constraints; and it does not incorporate the preferences of individual students (or teachers) for specific teachers (or students).

How does the software assign students to class-sections when there are multiple sections of a course on offer within the same period of the school day? The software's default priority is to balance across sections, meaning that when students are loaded into a course with multiple sections, the scheduler "always attempts to keep an even number of students among sections"(PowerSchool, 2026c,b). And while public documentation does not reveal the precise search routine or tie-breaking rule, published research in the operations literature on school timetabling nearly always features stochastic elements, frequently in the form random student ordering (e.g. Ceschia et al., 2023; Pillay, 2014; de Haan et al., 2006; Müller, 2005; Müller and Murray, 2010; Murray and Müller, 2007). Moreover, we will present evidence in section 3.2 that last-mile assignments made at this level are orthogonal to observed student characteristics. This is consistent with the idea that these "last-mile" assignments among students enrolled in the same academic course, in the same period, in the same school, in the same year, may reflect randomized ordering that effectively functions as lottery-based tie-breaking—in effect, creating a randomized controlled trial that we can use to assess the influence of classroom and teacher assignment.

However, a key threat to identification remains: the potential for an "aftermakerket" to invalidate this design by allowing for selection after course assignments are made. For example, it could be that some parents (e.g. those that are well-resourced) ask district administrators for manual changes after the software has made the initial assignment. There are at least three reasons this problem is not a concern in practice. First, accomodating student/parental requests in the aftermarket creates clear moral hazard problems and hence districts strongly discourage this practice. For example, CFISD's Cypress Creek High School states that schedules will not be changed to accommodate "teacher/lunch change requests" and that post-deadline changes are limited to level changes, error corrections, and adjustments generated by master-schedule changes (Cypress Creek High School, 2026). Second, and most importantly, the data we will use to measure student schedules are comprised of the initial

course assignments submitted to the state and hence less subject to manipulation.[3] Third and finally, we expect that the existence of an aftermarket would lead to testable forms of imbalance—for example, if higher-SES, more well-resourced parents are able to convince district administrators to place their child into sections with better, more experienced teachers, then we would expect to find that indicators of economic disadvantage are correlated with teacher experience even within "last-mile" course-period-year assignment cells. In practice, and as we will show in the next section, baseline and otherwise exogenous student characteristics are balanced which should mitigate this concern.

3.2 Empirical strategy

Motivated by the institutional details described in the preceding section, our research design seeks to isolate "last-mile," variation in classroom assignment. Our identifying assumption is that, conditional on taking the same course in the same period in the same school in the same year, the parallel classroom a student is placed into is driven by the constraints of the scheduling software and hence as-good-as random.

We now formalize that design. Let g index a last-mile assignment strata. Throughout, a stratum is defined as the complete interaction of course, period, and year within school which we will denote by the index g. Let $\mathcal{J}_g$ denote the set of classrooms in stratum g, and let D_{icg} equal one if student i is assigned to classroom $c \in \mathcal{C}_g$. For each feasible classroom c, define the potential outcome $Y_{ig}(c)$ that would be realized if student i were assigned to that classroom. The observed outcome is therefore

$$Y_{ig} = \sum_{c \in \mathcal{C}_g} D_{icg} Y_{ig}(c). \tag{1}$$

Our identifying assumption—that within a last-mile assignment stratum, classroom assignment is as-good-as-random—is formalized as:

$$D_{ig} \perp\!\!\!\perp \{Y_{ig}(c) : c \in \mathcal{C}_g\} \mid g, \tag{2}$$

Equation (2) is the formal counterpart to the institutional claim in the previous section: once attention is restricted to students taking the same course in the same period in the same year in the same school, the way the software assigns students across parallel classrooms is as-good-as random.[4]

The potential outcome model and identifying assumptions suggests the following regression:

[3] However, we acknowledge that the choice to use initial course assignment does come at a small cost: the potential for non-compliance if students switch out of their initially assigned class. For that reason, our current estimates are properly interpreted as Intention-to-treat (ITT) effects. In ongoing work, we are using course completion data to empirically measure how consequential non-compliance is for our design.

[4] This design is closest in spirit to the within-track high-school comparisons in Jackson (2014b), but we tighten the comparison to course × period × year cells in order explicitly isolate last-mile variation from the scheduling algorithm that is as-good-as random; by way of contrast, Jackson (2014b) conditions on a broader notion of "track" in order isolate variation in teacher quality that is orthogonal to curricula.

$$Y_{ig} = \alpha_g + \sum_{c \in \mathcal{C}_g \setminus \{c_0(g)\}} \delta_{cg} D_{icg} + \varepsilon_{ig}, \tag{3}$$

where α_g are last-mile-strata fixed effects and $c_0(g)$ denotes the omitted classroom in stratum g. In practice, we choose the omitted classroom at random within each stratum. Under (2), the coefficients δ_{cg} are causal within-stratum classroom effects relative to that omitted classroom.

Equation (3) makes clear that the treatment in our setting is classroom assignment itself. This is important substantively: assignment to a classroom bundles together the assigned teacher and any classroom-level inputs that move with that assignment, including realized peer composition and any residual variation in class size or room-specific resources. Our estimates from model (3) should therefore be interpreted as reduced-form effects of classroom assignment at the last mile.

For most of this extended abstract, however, a fully saturated classroom-effect model is not the most useful way to summarize the variation. The first issue is interpretation. By construction, each coefficient in (3) is defined relative to the omitted classroom in its own stratum, and that omitted teacher may be better or worse than the focal teacher. Randomly choosing the omitted classroom prevents the estimates from being mechanically anchored to a substantively chosen reference teacher, but it does not put coefficients from different strata on a common scale.

The second issue is statistical power. By statistical standards, classrooms are small. In the common case where a stratum contains two classrooms, (3) is numerically equivalent to a within-stratum difference in means between the focal classroom and the omitted classroom. Thus, each classroom contrast is often identified from only about 15–35 students in the focal classroom and 15–35 students in the reference classroom. The saturated specification therefore delivers a large collection of unbiased but noisy stratum-specific contrasts. A lower-dimensional representation can improve precision by pooling the components of classroom effects that are common across teachers and classrooms. We will therefore frequently find it convenient to summarize classroom assignment by decomposing these classroom effects into a component explained by teacher characteristics and a residual classroom component.

Let T_{cg} denote a teacher characteristic attached to classroom c in stratum g—for example, years of experience, years of education, gender, or race/ethnicity—and define $T_{ig} = \sum_{c \in \mathcal{C}_g} D_{icg} T_{cg}$ as the characteristic of the teacher assigned to student i. Thus, we will also be interested in estimating

$$Y_{ig} = \alpha_g + T'_{ig}\beta + u_{ig}. \tag{4}$$

Equation (4) can be viewed as the student-level projection of the saturated classroom effects in (3) onto the span of observed teacher characteristics. Equivalently, if the enrollment-weighted classroom effect in stratum g is written as

$$\delta_{cg} = T'_{cg}\beta + \omega_{cg}, \tag{5}$$

then (4) estimates the component of classroom assignment explained by T_{cg}, while ω_{cg} captures

residual classroom heterogeneity that is orthogonal to the observed teacher characteristics in the sense of a linear projection. This parameterization has two advantages. First, the coefficient β is on a common and easily interpretable scale across all strata. Second, it is substantially better powered than the saturated classroom-indicator model. Rather than estimating a separate effect for every classroom relative to a stratum-specific reference, (4) aggregates information across all strata about the covariance between outcomes and observed teacher attributes. In terms of (5), estimating β pools the signal in the high-dimensional vector $\{\delta_{cg}\}_{c,g}$ onto the low-dimensional span of $\{T_{cg}\}_{c,g}$.

A useful implication of (4) is that balance and identification can be assessed in the same low-dimensional language. In principle, one could test balance by regressing predetermined student characteristics on the full set of classroom indicators in (3). In practice, that approach inherits the same interpretation problem as the saturated treatment itself, since each classroom indicator is defined relative to a random omitted classroom. It is also poorly powered for exactly the same reason: the individual classroom contrasts are estimated from small cells and the implied balance tests are correspondingly noisy. We therefore test for balance using teacher characteristics, which are low-dimensional functions of classroom assignment and directly match the treatment variation emphasized in many of the main results.

Specifically, for each baseline student characteristic Z_{ig}, we estimate regressions of the form

$$Z_{ig} = \alpha_g + \beta T_{ig} + \varepsilon_{ig}, \tag{6}$$

where T_{ig} is a scalar teacher characteristic for the classroom assigned to student i, and g again indexes last-mile assignment strata. In practice, we estimate separate specifications using years of teacher experience, years of teacher education, an indicator for female teachers, and an indicator for non-white teachers. Baseline student characteristics include demographics (gender, race, age, economic status), education designations (english language learner, special education, at risk of not graduating, and gifted), and previous standardized test scores (Algebra I, Biology, English I, English II, 8th Grade Math, 8th Grade Reading).[5]

Figure 1 plots the coefficients from (6), normalized into statewide student-standard-deviation units to facilitate visual comparisons across outcomes with different scales. For each student characteristic, the figure also reports the corresponding *unconditional* estimate that omits the strata fixed effects α_g. The contrast is informative. Without α_g, the coefficients reflect the substantial selection that occurs across courses, periods, years, and schools. Once α_g is included, identification comes only from within-strata comparisons across classrooms that are as-good-as randomly assigned within school. In our data, the unconditional estimates reveal pronounced sorting on observables, whereas the within course-period-year estimates collapse toward zero and are jointly consistent with balance. Appendix Table A.1 reports the corresponding unstandardized coefficients and the p-value from the joint test across all 64 coefficients. This pattern is the empirical counterpart to the institutional detail discussed previously: selection is readily visible on the margins where students

[5] Only scores prior to the relevant class are included. Example: for a student in English I, only 8th grade math and 8th grade reading scores are treated as baseline characteristics.

choose courses and schedules, but it largely disappears on the final margin where the software allocates students across parallel classrooms.

Unless otherwise noted, all specifications report standard errors that are two-way clustered at the course-period-year stratum g and the student level. Clustering by g allows arbitrary dependence among students who are assigned within the same course-period-year—that is, the strata within which treatment is assigned. Clustering by student accounts for the fact that the same student can appear in more than one stratum. Identification therefore comes entirely from within-stratum comparisons across parallel classrooms, with inference that allows for both common assignment shocks within a stratum and repeated observations of the same student.

[Figure 1 about here.]

[Figure 2 about here.]

4 Preliminary Results

We emphasize at the outset that the estimates in this section should be interpreted as preliminary. Because disclosure was completed only for the Cypress-Fairbanks ISD sample at the time of this submission, all of the results below come from a single large district, even though nine additional districts are already merged to the ERC backbone and will enter subsequent drafts. In addition, the long-run outcomes are currently observed for fewer cohorts and fewer classroom cells than contemporaneous test scores, so the corresponding estimates are necessarily less precise. The value of the present evidence is therefore that, together with the balance results in Figure 1 and Appendix Table A.1, it shows that the research design is capable of recovering quasi-experimental variation in high-school classroom assignment, and that this assignment is capable of precisely estimating causal changes in outcomes even on the truncated sample.

4.1 Returns to Experience

We begin with teacher experience because it is has been heavily studied in prior observational work, generally using value-add type methods (e.g. Podolsky et al., 2019; Jackson et al., 2014; Rockoff, 2004; Rivkin et al., 2005; Clotfelter et al., 2007; Harris and Sass, 2011; Wiswall, 2013; Papay and Kraft, 2015; Ladd and Sorensen, 2017, among others) and hence it provides a benchmark by which to assess the power of our research design. More substantively, experience is interesting because it is the primary margin rewarded by public-school salary schedules. If the returns to experience are high, then rigid union salary structures that reward seniority could, in theory, be justified on the basis of it's ability to attract and retain more experienced teachers.

Figure 2 and Table 2 provide results from equation 4, but where the teacher characteristic of interest consists of mutually exclusive indicators for 1–5, 6–10, 11–15, and 16 or more years of experience (with first-year teachers as the omitted group). Since all specifications include the α_g fixed effects, the resulting experience profile is identified from comparisons of students who take the

same course in the same period in the same year in the same school, but that are quasi-randomly assigned across teachers with different experience levels.

[Figure 3 about here.]

[Table 2 about here.]

The test-score results reveal a clear and monotone return to experience as estimated within assignment strata. Relative to first-year teachers, assignment to a teacher with 1–5 years of experience raises test scores by 0.019σ (s.e. $= 0.011$), with effects rising to 0.025σ (0.013) for 6–10 years, 0.034σ (0.013) for 11–15 years, and 0.046σ (0.013) for 16 or more years; unsurprisingly, the joint test of equality to zero is rejected at $p = 0.005$.

Two features of this curve are worth emphasizing. First, the gradient is economically meaningful: the gap between first-year and 16+-year teachers is roughly two-fifths of the 0.12σ signal standard deviation of classroom effects on test scores will document below in section 4.3. Second, the profile shows little evidence of flattening within the experience range we observe. In the current data, experience continues to matter well into later-career bins—a contrast to prior observational work using value-added methods, which has general found the curve to be concave over this range (e.g. Rockoff, 2004).

The corresponding long-run estimates are much less precise. For bachelor's attainment, the point estimates are small and slightly negative relative to first-year teachers, ranging from -0.8 to -0.4 percentage points across the four experience bins, and they are jointly insignificant ($p = 0.731$). For log wages at age 25, the point estimates are positive but imprecise, ranging from 0.017 to 0.027 log points, with a joint p-value of 0.807. However, we do not interpret these nulls as evidence that experience matters only for short-run achievement. A more cautious reading is that the current single-district sample is powerful enough to detect the experience gradient in test scores but not yet large enough to pin down long-run effects of the corresponding magnitude. In fact, the point estimates themselves suggest that assignment to a non-novice teacher raises wages at age 25 by roughly 2.5 percentage points—a substantial earnings boost. Thus, one of the main benefits of the full multi-district expansion will be to determine whether the within-cell test-score gains associated with teacher experience translate into economically meaningful differences in adult outcomes.

4.2 Student-Teacher Match Effects

We next use the same within-assignment strata design to study whether some teachers are especially effective for certain types of students. Panel A of Figure 3 and Panel A of Table 3 estimate a baseline role-model specification with an indicator for assignment to a female teacher, an indicator for female students, and their interaction. The estimates reveal striking heterogeneity in contemporaneous achievement. For boys, assignment to a female teacher has essentially no effect on test scores (-0.007σ, s.e. $= 0.013$). For girls, the implied effect is approximately 0.092σ, obtained by adding the main female-teacher coefficient to the female-student interaction. Equivalently, the interaction term

itself is 0.099σ (0.015), a magnitude that exceeds the entire gap between first-year and 16+-year teachers documented in Figure 2.

[Figure 4 about here.]

[Table 3 about here.]

The longer-run outcomes are more nuanced and do not map neatly into a simple same-sex role-model story. In bachelor's attainment, female teachers raise boys' likelihood of attainment by a statistically significant 1.3 percentage points, while the female-student interaction is -1.3 percentage points, implying essentially zero net effect for girls. For log wages at age 25, the point estimates are small and imprecise for both sexes. The lesson is not that teacher-student matches are unimportant, but rather that the relevant dimensions of match quality are likely richer than a simple own-sex indicator, and that short-run academic complementarities need not translate mechanically into the same ranking of long-run outcomes.

Panel B of Figure 3 and Panel B of Table 3 repeat this exercise, but add interactions with an indicator for STEM courses. This is motivated by a large body of work hypothesizing that female under-representation in STEM fields is due to "pipeline" effects that emerge as a lack of female role models. We find that, outside STEM, female teachers increase girls' test scores by about 0.089σ and have slightly negative effects for boys (-0.025σ). In STEM classrooms, however, boys receive an additional 0.079σ boost from female teachers, while the STEM differential for girls is essentially zero once the triple interaction is accounted for. Thus the strong positive female-teacher effect for girls does not appear to be uniquely a STEM phenomenon; instead, female teachers seem to be especially effective for girls more generally, with an additional offsetting gain for boys specifically in STEM courses. The analogous heterogeneity in bachelor's attainment and wages is intriguing but too imprecisely estimated in the current sample to support strong conclusions. We therefore view Figure 3 primarily as proof of concept: the design has enough resolution to identify not only average teacher effects, but systematic heterogeneity in those effects by student characteristics and subject matter.

4.3 The Distribution of Classroom Effects

Teacher characteristics are useful because they deliver interpretable and relatively well-powered average effects, but they cannot capture the full heterogeneity revealed by the design. We therefore return to the saturated classroom-assignment specification and examine the cross-classroom distribution of reduced-form classroom effects. Here, a classroom effect should be read as a teacher-by-course effect: it measures the causal effect of assignment to a particular classroom relative to other classrooms in the same course-period-year cell, bundling the assigned teacher with any classroom-level inputs that move with that assignment.

Table 4 shows substantial dispersion in these effects, especially for test scores. The joint null that all classroom effects are zero is decisively rejected for test scores ($p < 0.001$), and the estimated

signal standard deviation is 0.12σ after correcting for measurement error in the raw cross-classroom dispersion. For log wages at age 25, the corresponding signal standard deviation is 0.082 log points—roughly 8 percent—and the joint test is marginally significant ($p = 0.076$), indicating suggestive heterogeneity in long-run labor-market effects even in the current single-district sample. By contrast, we detect essentially no signal in bachelor's attainment: the joint p-value is 0.988 and the estimated signal standard deviation is only 0.008. However, we note that this is likely due to excessive noise in the classroom estimates; as discussed in the preceding section, we find evidence of important causal effects for this outcome when we take the more powerful approach of projecting them onto teacher and student characteristics. Thus, we expect to sharpen these estimates as we continue to expand our sample.

[Table 4 about here.]

Figure 4 asks whether the classrooms that raise test scores are also the classrooms that improve later outcomes. The answer is yes. Across classroom cells, the estimated effect on bachelor's attainment rises with the estimated effect on test scores, with a raw OLS slope of 0.089 (0.014). The analogous slope for log wages at age 25 is 0.163 (0.047). However, we sound a note of caution on these results: they relate estimated classroom effects to one another and therefore inherit measurement error from both axes. That caveat matters, because classical error in the test-score effect mechanically attenuates the slope toward zero, while non-classical measurement error can cause changes that move in unpredictable directions. In ongoing work, we are estimating models on the micro-data that will explicitly correct for the bias caused by measurement error (both classical and non-classical). However to get a sense of how consequential they are for the point estimates we have currently disclosed, Appendix Table A.2 probes how these results vary with both classical and non-classical measurement error. We find classical measurement-error-adjusted slopes of 0.242 for bachelor's attainment and 0.350 for log wages. Moreover, the residual correlation across outcomes necessary for non-classical measurement error to "undo" this finding (in the sense of losing statistical significance) would have to be large ($\rho = 0.16$ for bachelor's attainment and $\rho = 0.1$ for log-wages), which we take as evidence that these results are likely robust to non-classical measurement error. At the same time, the modest R^2 values in Figure 4 remind us that test-score impacts do not exhaust all classroom heterogeneity. Test scores appear to capture an important component of broader classroom quality, but not the entirety of it.

[Figure 5 about here.]

The economic stakes implied by this heterogeneity are large. Figure 5 converts a one-standard-deviation improvement in classroom quality into the present discounted value of lifetime earnings gains for a single teacher-year of students. Depending on whether classroom quality is measured directly from the distribution of log-wage effects or indirectly via the test-score-to-wage relationship, and depending on assumptions about class load and discounting, the implied gain to a 1-SD improvement in classroom quality ranges from roughly \$2.38 million to \$10.54 million, with our

preferred estimate landing at $3.52 million. These calculations are deliberately back-of-the-envelope, and they should not be read as welfare estimates. But they make a simple point that survives a wide range of assumptions: because high-school teachers teach many students each year, even modest improvements in classroom quality can generate very large aggregate gains.

[Figure 6 about here.]

5 Conclusion: Ongoing and Future Work

The results in this extended abstract should be interpreted as a proof of concept for the research design rather than as the final word on the economic and social value of high-school teachers. Using the currently disclosed results from a single-district sample, we document balance on a rich set of baseline student characteristics and show that the design is already informative for both short-run achievement and selected long-run outcomes. Even in this restricted sample, the estimates suggest several substantive patterns worthy of further investigation: returns to experience appear clearly increasing for test scores but less obviously so for wages; match effects are present but need not align across contemporaneous and long-run outcomes; and the dispersion in classroom effects is wide enough to imply economically meaningful differences in adult earnings. With that said, we are reluctant to over-interpret those magnitudes on the basis of one district. But taken together, the current results show that the design has the potential to illuminate dimensions of educational production that have been difficult to study using conventional value-added type methods. The next phase of the project is therefore less about changing the design than about scaling it—across districts, across outcomes, and from reduced-form classroom effects to persistent teacher effects. All three steps are already underway, and we expect to have complete, polished versions in time for the Summer Institute. Thus, we conclude by discussing these steps in more detail.

Additional data. The principal limitation of the current draft is disclosure, not data construction. Since producing the estimates reported here, we have merged course schedule files for nine additional Texas districts into the ERC backbone (which was itself a time consuming process), bringing the working database to approximately 1,000,000 student $\times$ course observations comprised of over 260,000 students and almost 6,000 teachers across ten districts. Preliminary versions of the balance (not yet disclosed) in these newly merged districts are promising: within-school, within-course, within-period assignment continues to look well balanced on observed student characteristics. That portability is important because it suggests that the quasi-experimental margin we exploit is not unique to Cypress-Fairbanks ISD, but instead reflects a more general feature of how large Texas high schools operationalize course scheduling. The district course files, and the research design they enable, are a key contribution of the paper. These records are not delivered as part of the standard ERC extract, and assembling them has required district-by-district public-information requests (116 submitted as of this writing), substantial harmonization of heterogeneous local formats, and then merging back to the ERC student and teacher records. Subsequent drafts will incorporate the

remaining nine districts as well as any additional districts we are able to secure and harmonize this spring. We are also actively submitting additional public-information requests to other districts and, as of this writing, have received data from an additional 21 districts; however, it remains to be seen whether we can successfully harmonize and merge them into the ERC backbone.

Additional outcomes. While our current proof-of-concept focusses on test scores, bacherlor's degree attainment, and earnings, our intention is to fully trace out students' educational experiences over the short, medium, and long-run. Additional schooling outcomes include attendance, discipline, election into future advanced coursework such as AP or honors classes, and college major choice.

In addition to more schooling outcomes, we have also expanded the set of post-education adult outcomes available for analysis. On March 20, 2026, TEA returned an anonymized merge initiated by us between the ERC and the L2 voter files. This is consequential for the project because it allows us to move beyond educational attainment and labor-market success to broader notions of adult well-being, including measures of civic participation such as voter registration and turnout. That extension is especially valuable in a paper about high-school teachers. Teachers may influence not only academic skill formation and later earnings, but also habits of participation, attachment to local institutions, and the broader set of behaviors associated with civic life. The L2 merge therefore opens a new empirical margin on which to evaluate teachers' long-run social value. More speculatively, we are investigating whether criminal-justice records that we believe have been linked into the Texas ERC for other projects can be made available for this study. At present, we do not yet know the relevant coverage, time span, or disclosure constraints, so we view crime outcomes as exploratory rather than committed for the July version of the paper. But if feasible, such records would further broaden the set of adult outcomes we can use to explore how high-school teachers shape the lives of the children they are charged with educating.

Deconvolving the distribution of teacher effects from classroom effects. A final priority for us is to move from the reduced-form distribution of classroom effects to the distribution of persistent teacher effects. The classroom estimates reported in this draft are causally identified by design, but they combine stable teacher quality with classroom-specific shocks (e.g. idiosyncratic peer composition), and traditional sampling variability. A useful conceptual representation is

$$\delta_{cg} = \tau_{j(c,g)} + \kappa_{cg} + \eta_{cg}, \tag{7}$$

where δ_{cg} is the causal effect of classroom c contained in randomization strata g, $\tau_{j(c,g)}$ is a persistent teacher component that we expect to span classes within and across a given academic year; κ_{cg} captures a classroom-specific component of the variance that does not disappear asymptotically as the number of students increases; and η_{cg} is a transitory shock (i.e. traditional sampling variability).

Our setting is unusually well suited to separately identifying these objects because we observe many teachers teaching multiple classrooms within the same year and repeatedly over time. That repeated structure allows us to adapt the logic of Kane and Staiger (2008) and Chetty et al.

(2014a,c) to a setting where we can observe long-run outcomes and where the underlying source of identification is design-based, quasi-experimental variation. Put differently, each teacher generates multiple quasi-experimental estimates of classroom impacts, and the covariance of those estimates across classes and cohorts can be used to recover the variance of persistent teacher effects net of sampling noise and classroom shocks. This was a central limitation of Project STAR, where researchers could identify causal classroom effects but did not observe the same teacher repeatedly across classrooms and hence could not distinguish teachers separately from the classroom component of the variance (Chetty et al., 2011). And while the MET project did have the repeated observations necessary to distinguish teacher effects from the classroom component, it has not yet (to our knowledge) been linked to long-run outcomes. The only paper that has come close to this goal is Chetty et al. (2014b); however, their empirical strategy is an event study design which was not equipped to analyze the full distribution of teacher effects, instead focusing on how teacher value-added translates into earnings on average.

In our setting we will, for the first time, be able to estimate the full distribution of teacher effects on earnings using a design-based source of quasi-random variation for identification. More broadly, this approach will let us quantify how much of the economically important dispersion documented in this draft reflects stable differences across teachers versus classroom-specific shocks, a distinction that matters directly for teacher policy and for the interpretation of value-added measures.

References

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics*, 25(1):95–135.

Backes, B., Cowan, J., Goldhaber, D., and Theobald, R. (2023). How to measure a teacher: The influence of test and nontest value-added on long-run student outcomes. Working Paper 270-0423-2, National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Ceschia, S., Di Gaspero, L., and Schaerf, A. (2023). Educational timetabling: Problems, benchmarks, and state-of-the-art results. *European Journal of Operational Research*, 308(1):1–18.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *Quarterly Journal of Economics*, 126(4):1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–2679.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014c). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–2679.

Clear Creek Independent School District (2026). Course Selection. District website. Accessed March 24, 2026.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6):673–682.

Cook, J. B. and Mansfield, R. K. (2016). Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers. *Journal of Public Economics*, 140:51–72.

Cy-Fair High School (2026). 2025–26 Course Information. Cypress-Fairbanks Independent School District website. Accessed March 24, 2026.

Cypress Creek High School (2026). Frequently Asked Questions. Cypress-Fairbanks Independent School District website. Accessed March 24, 2026.

de Haan, P., Landman, R., Post, G., and Ruizenaar, H. (2006). A four-phase approach to a timetabling problem in secondary schools. In *Proceedings of the 6th International Conference on the Practice and Theory of Automated Timetabling (PATAT 2006)*, pages 423–425, Brno, Czech Republic.

Gilraine, M., Gu, J., and McMillan, R. (2020). A new method for estimating teacher value-added. NBER Working Paper 27094, National Bureau of Economic Research.

Gilraine, M. and Pope, N. G. (2021). Making teaching last: Long-run value-added. NBER Working Paper 29555, National Bureau of Economic Research.

Goldhaber, D. D., Goldschmidt, P., and Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis*, 35(2):220–236.

Harris, D. N. and Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8):798–812.

Jackson, C. K. (2012). Do high school teachers really matter? Working Paper WP-12-09, Institute for Policy Research, Northwestern University.

Jackson, C. K. (2014a). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, 32(4):645–684.

Jackson, C. K. (2014b). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, 32(4):645–684.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, 6(1):801–825.

Jordan High School (2026). Course Selection. Katy Independent School District website. Accessed March 24, 2026.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Working Paper 14607, National Bureau of Economic Research.

Koedel, C. and Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47:180–195.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2):497–532.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *Economic Journal*, 111(468):1–28.

Ladd, H. F. and Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy*, 12(2):241–279.

Lavy, V. and Megalokonomou, R. (2024). Alternative measures of teachers' value added and impact on short and long-term outcomes: Evidence from random assignment. NBER Working Paper 32671, National Bureau of Economic Research.

Liu, J. and Loeb, S. (2021). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, 56(2):343–379.

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, 33(3):751–788.

Mulhern, C. and Opper, I. (2021). Measuring and summarizing the multiple dimensions of teacher effectiveness.

Müller, T. (2005). *Constraint-based Timetabling*. PhD thesis, Charles University in Prague, Faculty of Mathematics and Physics, Prague.

Müller, T. and Murray, K. (2010). Comprehensive approach to student sectioning. *Annals of Operations Research*, 181(1):249–269.

Murray, K. and Müller, T. (2007). Real-time student sectioning. In *Proceedings of the 3rd Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA 2007)*, pages 598–600, Paris, France. Conference abstract.

National Center for Education Statistics (2024). Public school enrollment. https://nces.ed.gov/programs/coe/indicator/cga/public-school-enrollment. Reports approximately 15.5 million students enrolled in grades 9–12 in fall 2022. Accessed March 2026.

Papay, J. P. and Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, 130:105–119.

Pillay, N. (2014). A survey of school timetabling research. *Annals of Operations Research*, 218(1):261–293.

Podolsky, A., Kini, T., and Darling-Hammond, L. (2019). Does teaching experience increase teacher effectiveness? a review of US research. *Journal of Professional Capital and Community*, 4(4):286–308.

PowerSchool (2026a). PowerScheduler. PowerSchool SIS Administrator Help, version 25.1. Accessed March 24, 2026.

PowerSchool (2026b). Run a Load. PowerSchool SIS Administrator Help, version 25.1. Accessed March 24, 2026.

PowerSchool (2026c). Verify and Complete Course Information. PowerSchool SIS Administrator Help, version 25.1. Accessed March 24, 2026.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2):247–252.

Rose, E. K., Schellenberg, J. T., and Shem-Tov, Y. (2025). The effects of teacher quality on adult criminal justice contact. Latest public draft.

Shah, H. (2025). High school teachers and college-relevant skill production: Evidence from act score value added. Job market paper.

Spillane Middle School (2026). FAQs for HS 4-Year Plan. Cypress-Fairbanks Independent School District website. Accessed March 24, 2026.

Tan, J. S., Goh, S. L., Kendall, G., and Sabar, N. R. (2021). A survey of the state-of-the-art of optimisation methodologies in school timetabling problems. *Expert Systems with Applications*, 165:113943.

Tartova, D. (2023). Teacher value-added in the absence of annual test scores: Utilizing teacher networks. SSRN working paper; revised February 2025.

Texas Education Agency (2026). State Graduation Requirements. TEA website. Accessed March 24, 2026.

Texas Education Research Center (2026a). History and Background. University of Texas at Austin.

Texas Education Research Center (2026b). Quick Overview of Data. University of Texas at Austin.

Texas Education Research Center (2026c). Type of Data. University of Texas at Austin.

Texas Schools Project (2026). Data Holdings. University of Texas at Dallas Education Research Center.

The Texas Tribune (2026a). About the Data. Texas Public Schools Explorer.

The Texas Tribune (2026b). Cypress-Fairbanks ISD. Texas Public Schools Explorer.

The Texas Tribune (2026c). State of Texas. Texas Public Schools Explorer.

U.S. Bureau of Labor Statistics (2023). Employment and wages for secondary school teachers. https://www.bls.gov/oes/current/oes252031.htm. Reports approximately 1.1 million secondary school teachers in the United States. Accessed March 2026.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, 100:61–78.

Figure 1: Balance on baseline student characteristics

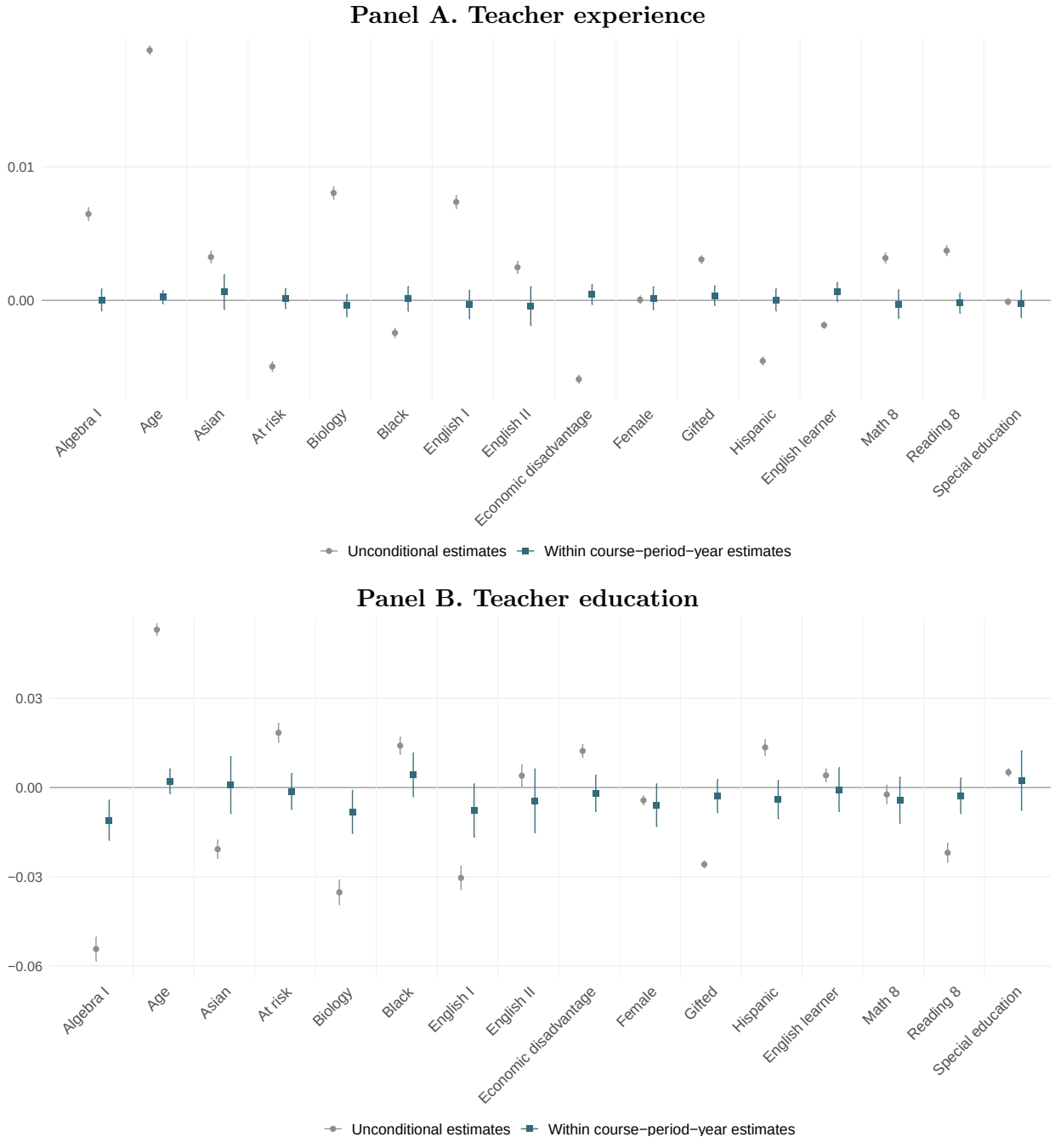
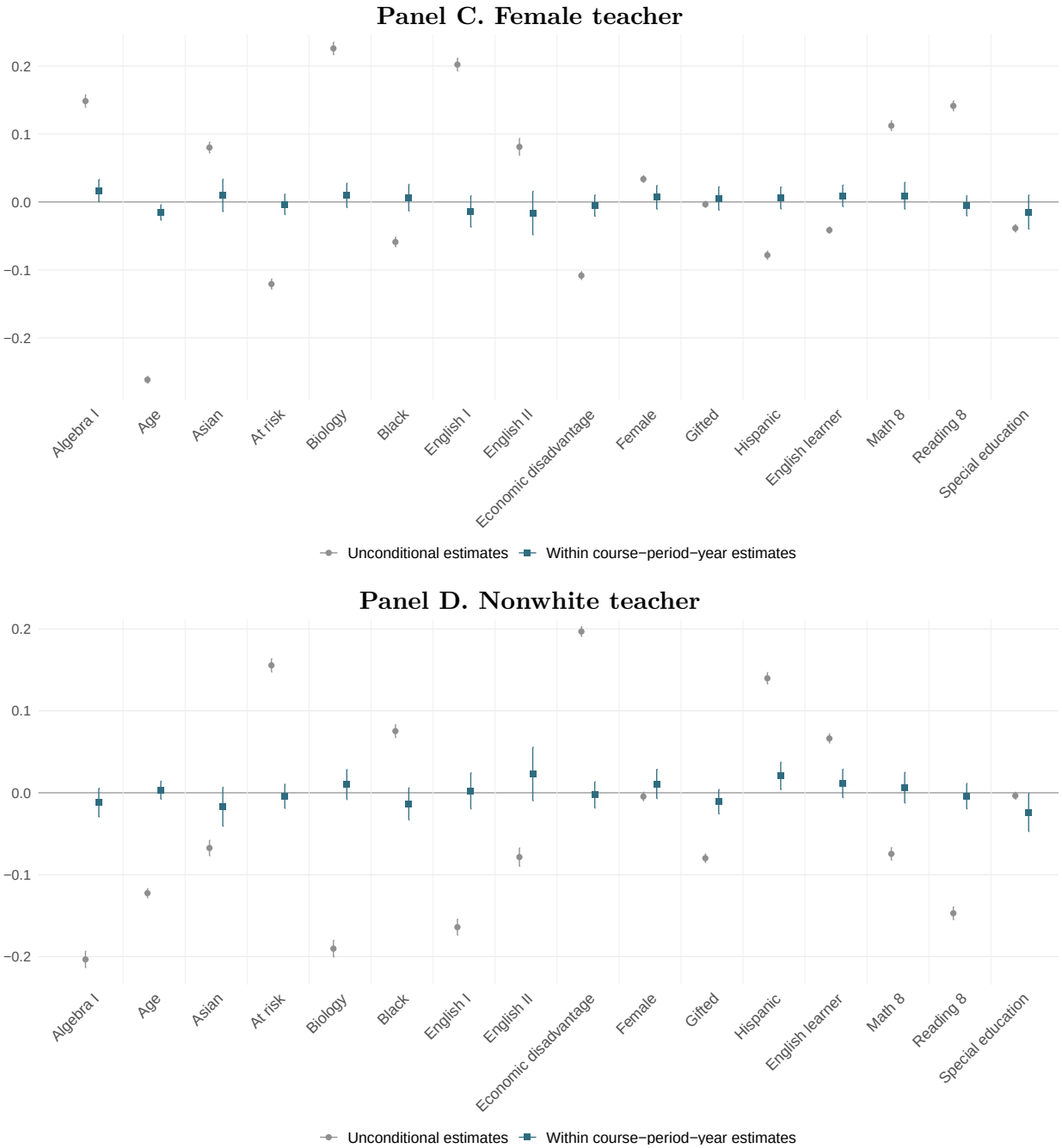
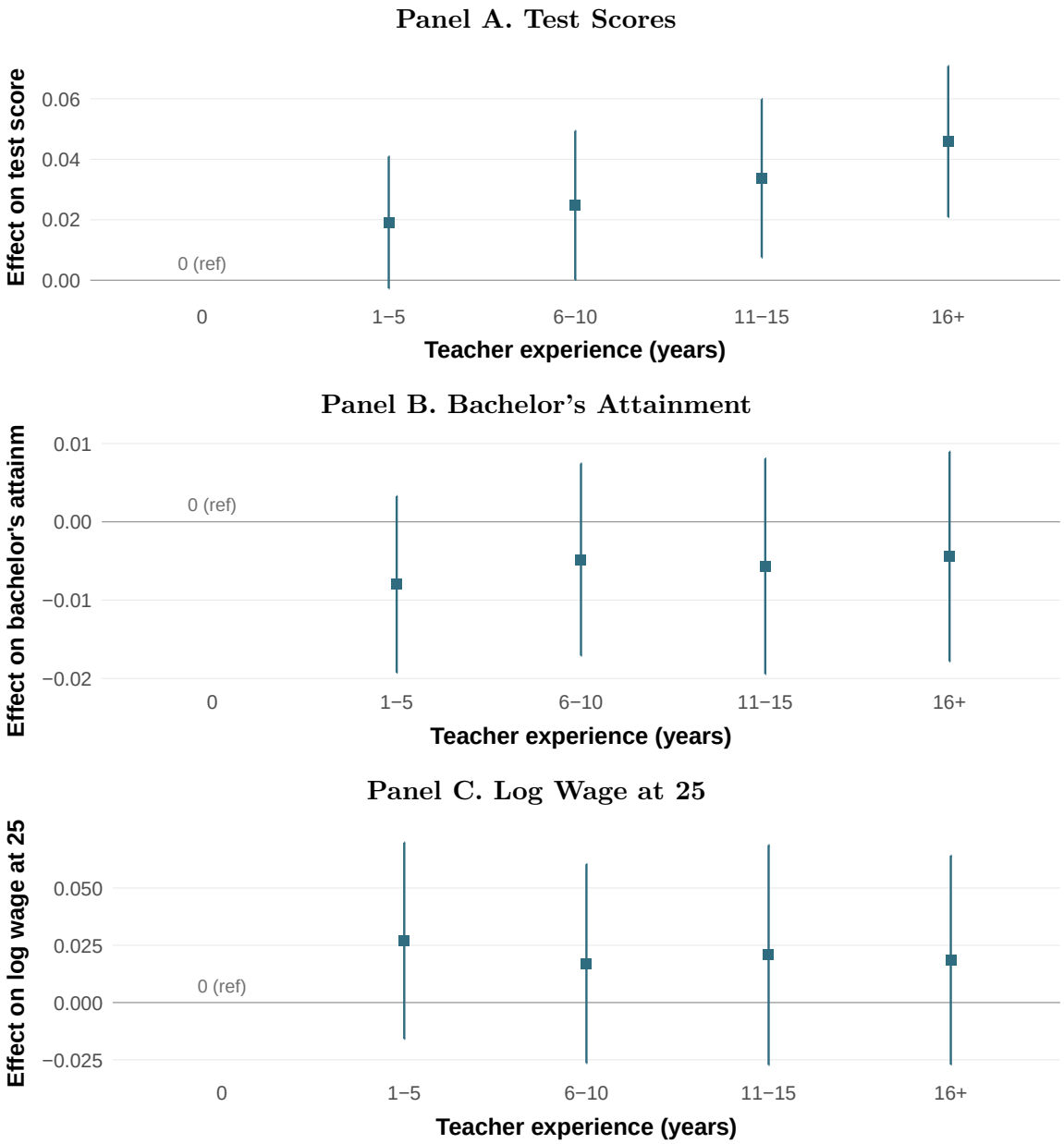


Figure 1: Balance on baseline student characteristics (continued)



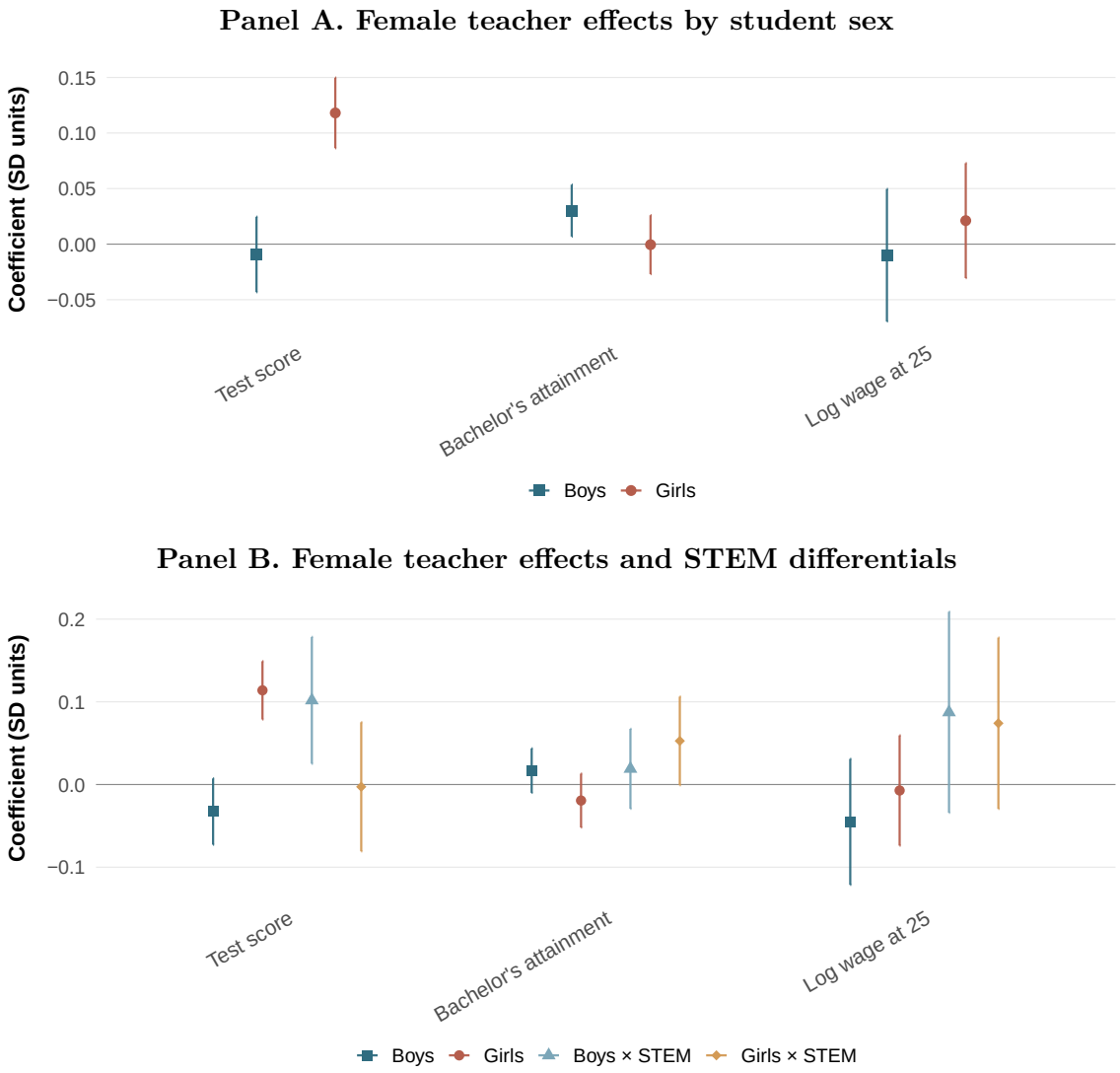
Notes: This figure plots coefficients from regressions of the form $Z_{ig} = \alpha_g + \beta T_{ig} + \varepsilon_{ig}$, where Z_{ig} is a baseline characteristic for student i, T_{ig} is a characteristic of the teacher assigned to student i, and g indexes "last-mile" assignment strata, defined as the complete interaction of course, period, and year. Panel A uses years of teacher experience for T_{ig}; Panel B uses years of teacher education; Panel C uses an indicator for female teachers; and Panel D uses an indicator for non-white teachers. The x-axis indexes alternative student characteristics Z_{ig}, and the y-axis reports the corresponding coefficients, normalized into statewide student-standard-deviation units to facilitate comparison across outcomes with different scales. The series labeled *Unconditional estimates* omit the strata fixed effects α_g and therefore reflect both selection and causal effects. The series labeled *Within course-period-year estimates* include α_g, so identification comes from within-strata comparisons across classrooms that are as-good-as randomly assigned. Standard errors are two-way clustered at the course-period-year (the strata within which "treatment" is assigned) and student (to account for the fact that students may appear in more than one strata) levels. Whiskers denote 95 percent confidence intervals. Unstandardized coefficients and a joint test of balance across all student covariates and teacher characteristics are reported in Appendix Table A.1.

Figure 2: Teacher experience and student outcomes



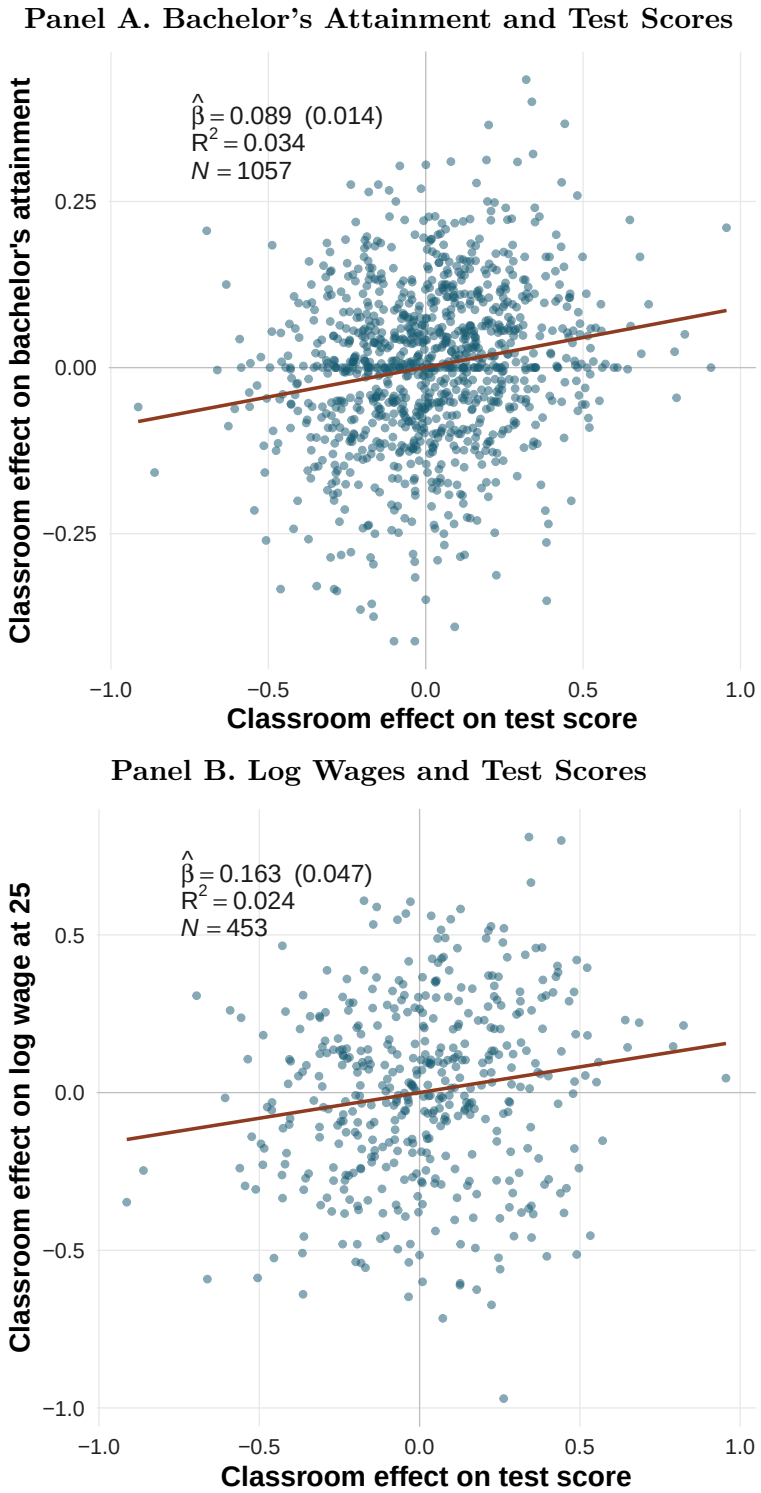
Notes: This figure plots coefficients from regressions of the form $Y_{ig} = \alpha_g + \beta_1 \mathbf{1}_{E_{ig} \in 1\text{–}5} + \beta_2 \mathbf{1}_{E_{ig} \in 6\text{–}10} + \beta_3 \mathbf{1}_{E_{ig} \in 11\text{–}15} + \beta_4 \mathbf{1}_{E_{ig} \geq 16} + \varepsilon_{ig}$, where Y_{ig} is the indicated outcome for student i, E_{ig} is the years of experience of the teacher assigned to student i, and g indexes last-mile assignment strata, defined as the complete interaction of course, period, and year. The omitted category is first year teachers. All specifications include last-mile assignment-strata fixed effects, so identification comes from within-strata comparisons across classrooms that are as-good-as randomly assigned. Standard errors are two-way clustered at the course-period-year and student levels, and whiskers denote 95 percent confidence intervals. The corresponding estimates are reported in table 2.

Figure 3: Teacher matching effects by student sex and STEM status



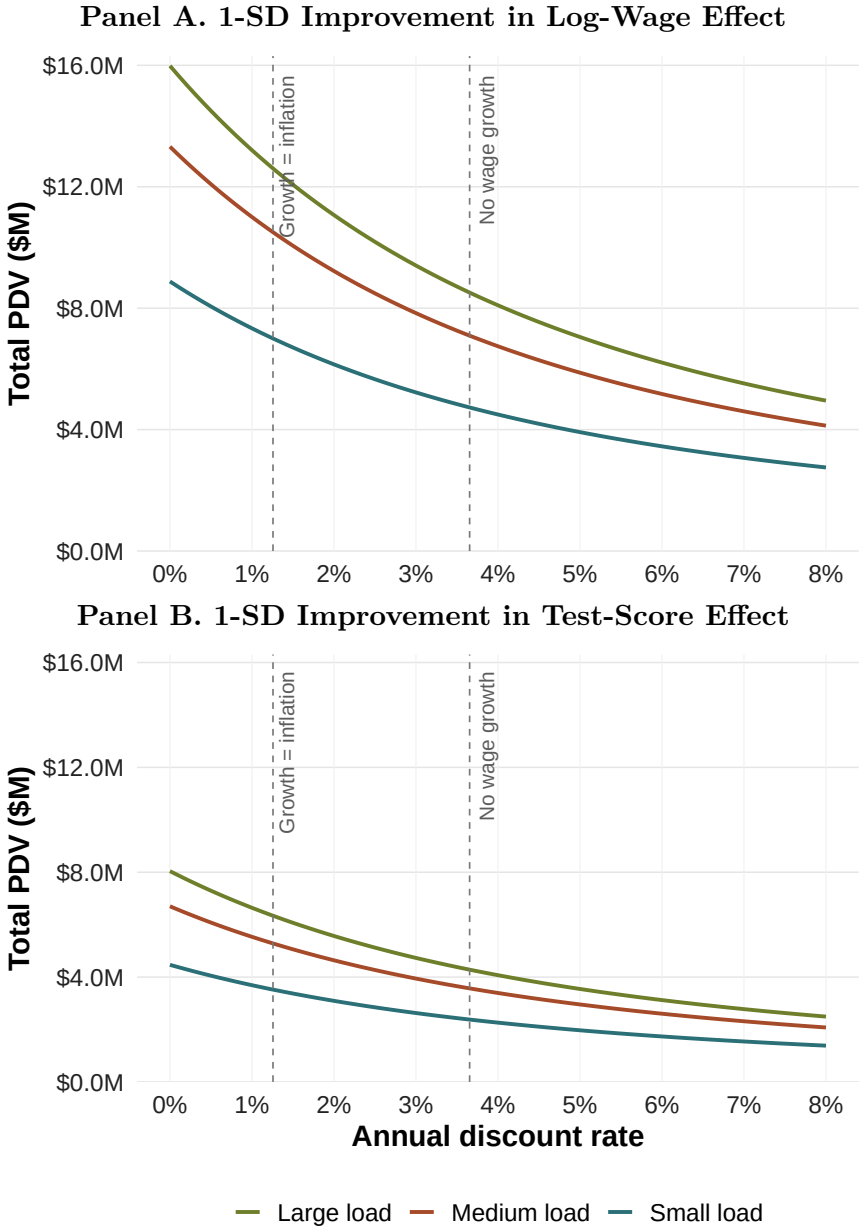
Notes: Panel A plots implied female-teacher effects from regressions of the form $Y_{ig} = \alpha_g + \beta_1 FTeach_{ig} + \beta_2 Female_i + \beta_3(FTeach_{ig} \times Female_i) + \varepsilon_{ig}$. The boys series reports β_1, while the girls series reports $\beta_1 + \beta_3$. Panel B is identical, except the model is fully saturated in an indicator for whether the course being taught is classified as STEM. All specifications include last-mile assignment-strata fixed effects, so identification comes from within-strata comparisons across classrooms that are as-good-as randomly assigned within a school. Coefficients are normalized into state wide outcome standard deviation units to facilitate visual comparability across outcomes on different scales. Whiskers denote 95 percent confidence intervals. Table 3 provides corresponding, unstandardized, point estimates.

Figure 4: Relationships between classroom effects in test scores and later outcomes



Notes: This figure plots classroom-level causal-effect estimates for later outcomes against classroom-level causal-effect estimates for test scores. Panel A plots classroom effects on bachelor's attainment against classroom effects on test scores. Panel B plots classroom effects on log wages at age 25 against classroom effects on test scores. Each point is a classroom-effect estimate from model 3. The fitted line is the raw OLS regression of the y-axis classroom effect on the x-axis classroom effect. The metadata report the slope estimate $\hat{\beta}$, its heteroskedasticity-robust standard error, the regression R^2, and the number of classroom cells used in the panel. These figures do not adjust for measurement error; Appendix Table A.2 reports slopes that have been adjusted for classical measurement error and a sensitivity analysis designed to assess the potential for correlated measurement error across outcomes to impact these results.

Figure 5: Aggregate lifetime earnings gains from a 1-SD improvement in classroom quality



Notes: This figure reports a back-of-the-envelope present discounted value (PDV) calculation for the total lifetime earnings gains generated by a one-standard-deviation improvement in classroom quality for a single teacher-year of students. Panel A measures classroom quality directly using the distribution of classroom effects on log wages. Panel B defines classroom quality using the distribution of classroom effects on test scores and maps that improvement into log wages using the classical measurement-error-adjusted relationship between classroom effects on log wages and classroom effects on test scores. In both panels, the implied annual earnings gain per student is computed as $\bar{w}_{25}(e^{\Delta \log w} - 1)$ and is assumed to persist from age 25 through age 64. The plotted PDV at discount rate r is $N_T \times \bar{w}_{25}(e^{\Delta \log w} - 1) \times \sum_{a=1}^{40} \psi_a (1+r)^{-a}$, where N_T is students taught per teacher-year and ψ_a is the survival probability at age $25 + a$ from the Social Security Administration actuarial life tables. The three colored lines correspond to small, medium, and large teacher-load assumptions: 5 classes of 20 students, 6 classes of 25 students, and 6 classes of 30 students. The dashed vertical lines mark the two benchmark discounting assumptions used in Appendix Table A.3: a growth-equals-inflation benchmark represented by a net discount rate of 1.26% and a no-wage-growth benchmark represented by a discount rate of 3.65%.

Table 1: Summary Statistics

	All Texas	CyFair District
Panel A. Student Demographic Characteristics		
Students	5,517,464	118,187
Black students	12.8%	20.6%
Hispanic students	53.2%	45.5%
Asian students	5.4%	10.0%
Economically disadvantaged	62.3%	60.1%
Limited English proficiency	24.4%	20.7%
Panel B. Student Outcome Variables		
Test score (standardized)	—	0.519σ
Bachelor's attainment	—	25.8%
Log wage at 25	—	10.56
Wage at 25 (winsorized)	—	28,400
Panel C. Teacher Characteristics		
Average teacher experience (years)	11.1	10.4
Bachelor's degree	71.7%	69.6%
Master's degree	24.9%	28.0%
Doctorate	0.8%	0.9%

Notes: This table combines statistics released through the Texas ERC with public statewide and district reference statistics. Panel A and Panel C use 2023–2024 Texas Public Schools Explorer statistics for the State of Texas and Cypress-Fairbanks ISD. Panel B reports means from the outcome variables used in the analysis. Missing entries (and our reliance on public sources for panels A and C) reflect the fact that many of these means were not disclosed in time for the Summer Institute submission deadline.

Table 2: Teacher experience and student outcomes

	Test score	Bachelor's attainment	Log wage at 25
1-5 years	0.019*	-0.008	0.027
	(0.011)	(0.006)	(0.022)
6-10 years	0.025**	-0.005	0.017
	(0.013)	(0.006)	(0.022)
11-15 years	0.034**	-0.006	0.021
	(0.013)	(0.007)	(0.024)
16+ years	0.046***	-0.004	0.019
	(0.013)	(0.007)	(0.023)
Observations	74,388	80,362	20,707
Joint p-value ($H_0 : \beta_1 = \cdots = \beta_4 = 0$)	0.005	0.731	0.807

Notes: This table reports coefficients from regressions of the form $Y_{ig} = \alpha_g + \beta_1 \mathbf{1}_{E_{ig} \in 1\text{–}5} + \beta_2 \mathbf{1}_{E_{ig} \in 6\text{–}10} + \beta_3 \mathbf{1}_{E_{ig} \in 11\text{–}15} + \beta_4 \mathbf{1}_{E_{ig} \geq 16} + \varepsilon_{ig}$, where Y_{ig} is the outcome for student i, E_{ig} is the years of experience of the teacher assigned to student i, and g indexes last-mile assignment strata, defined as the complete interaction of course, period, and year. The omitted category is teachers with 0 years of experience. The columns report contemporaneous test scores, bachelor's attainment, and log wages at age 25. All specifications include last-mile assignment-strata fixed effects, so identification comes from within-strata comparisons across classrooms that are as-good-as randomly assigned. Standard errors are two-way clustered at the course-period-year and student levels and are reported in parentheses. The bottom joint p-value tests the null that all four non-omitted experience-bucket coefficients are zero. Significance stars denote $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$. Figure 2 provides a visualization of these coefficients.

Table 3: Teacher matching regressions

	Test score	Bachelor's attainment	Log wage at 25
Panel A. Baseline role-model specification			
Female teacher	-0.007	0.013**	-0.006
	(0.013)	(0.005)	(0.020)
Female teacher × female student	0.099***	-0.013**	0.020
	(0.015)	(0.007)	(0.022)
Observations	74,388	80,362	20,707
Panel B. Role-model specification with STEM interactions			
Female teacher	-0.025	0.007	-0.029
	(0.016)	(0.006)	(0.025)
Female teacher × female student	0.114***	-0.016**	0.024
	(0.018)	(0.008)	(0.026)
Female teacher × STEM	0.079***	0.008	0.056
	(0.030)	(0.011)	(0.040)
Female teacher × female student × STEM	-0.081**	0.015	-0.009
	(0.036)	(0.013)	(0.045)
Observations	74,388	80,362	20,707

Notes: Each column reports a separate outcome regression. Panel A plots implied female-teacher effects from regressions of the form $Y_{ig} = \alpha_g + \beta_1 FTeach_{ig} + \beta_2 Female_i + \beta_3 (FTeach_{ig} \times Female_i) + \varepsilon_{ig}$. Panel B is identical, except the model is fully saturated in an indicator for whether the course being taught is classified as STEM. All models include last-mile assignment-strata fixed effects, so identification comes from within-strata comparisons across classrooms that are as-good-as randomly assigned. Standard errors are two-way clustered at the course-period-year and student levels and are reported in parentheses. Figure 3 provides a corresponding visualization of these coefficients. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Distribution of classroom effects across outcomes

Outcome	Joint-null p-value	Signal σ
Test score	< 0.001	0.12
Bachelor's attainment	0.988	0.00789
Log wage at 25	0.076	0.0821

Notes: Each row summarizes the estimated distribution of classroom (i.e. teacher-by-course) effects for one outcome. The first column reports the p-value from the joint null test that all teacher-course effects are zero. The second column reports the estimated signal standard deviation (i.e. an estimate of $\sqrt{\text{Var}(\delta_c)}$). Let $\hat{\delta}_c$ denote the estimated teacher-course effect for classroom c, and write $\hat{\delta}_c = \delta_c + u_c$. Let $\widehat{SE}_c$ denote the reported standard error of $\hat{\delta}_c$. Then the estimate of the signal variance is given by $\hat{\sigma}^2 = \text{Var}_c(\hat{\delta}_c) - \frac{1}{C}\sum_{c=1}^{C} \widehat{SE}_c^2$, that is, the cross-classroom dispersion net of average estimated sampling variance. For ease of interpretation, we take the square root and report the signal standard deviation.

Appendix

A Additional Tables and Figures

Table A.1: Balance on baseline student characteristics

	Assigned teacher characteristic			
Student covariate	Experience	Female	Nonwhite	Education
Algebra I	0.00003	0.01393**	-0.01003	-0.00916***
	(0.00035)	(0.00691)	(0.00740)	(0.00288)
Age	0.00024	-0.01559***	0.00316	0.00213
	(0.00025)	(0.00578)	(0.00563)	(0.00214)
Asian	0.00013	0.00203	-0.00360	0.00018
	(0.00014)	(0.00255)	(0.00254)	(0.00103)
At risk	0.00007	-0.00177	-0.00215	-0.00064
	(0.00019)	(0.00381)	(0.00377)	(0.00154)
Biology	-0.00033	0.00826	0.00856	-0.00699**
	(0.00036)	(0.00780)	(0.00796)	(0.00315)
Black	0.00004	0.00229	-0.00491	0.00156
	(0.00017)	(0.00362)	(0.00362)	(0.00136)
English I	-0.00025	-0.01106	0.00189	-0.00609*
	(0.00043)	(0.00928)	(0.00884)	(0.00361)
English II	-0.00033	-0.01273	0.01789	-0.00349
	(0.00058)	(0.01275)	(0.01300)	(0.00428)
Economic disadvantage	0.00022	-0.00271	-0.00131	-0.00098
	(0.00019)	(0.00400)	(0.00407)	(0.00156)
Female	0.00008	0.00341	0.00539	-0.00295
	(0.00022)	(0.00442)	(0.00451)	(0.00183)
Gifted	0.00008	0.00126	-0.00265	-0.00070
	(0.00009)	(0.00211)	(0.00182)	(0.00069)
Hispanic	0.00002	0.00299	0.01029**	-0.00204
	(0.00021)	(0.00411)	(0.00425)	(0.00165)
English learner	0.00019*	0.00273	0.00346	-0.00021
	(0.00011)	(0.00245)	(0.00268)	(0.00114)
Math 8	-0.00027	0.00862	0.00582	-0.00402
	(0.00051)	(0.00941)	(0.00893)	(0.00371)
Reading 8	-0.00020	-0.00521	-0.00391	-0.00261
	(0.00036)	(0.00709)	(0.00739)	(0.00286)
Special education	-0.00006	-0.00342	-0.00553**	0.00054
	(0.00012)	(0.00293)	(0.00272)	(0.00117)

Joint test of balance across all columns and covariates
$P(F < f \mid H_0 : \beta_1 = \cdots = \beta_{64} = 0) = 0.148$

Notes: This table reports coefficients from regressions of the form $Z_{ig} = \alpha_g + \beta T_{ig} + \varepsilon_{ig}$, where Z_{ig} is a baseline characteristic for student i, T_{ig} is a characteristic of the teacher assigned to student i, and g indexes last-mile assignment strata, defined as the complete interaction of course, period, and year. The estimates include last-mile-strata fixed effects α_g, so identification comes from within-strata comparisons across classrooms that are as-good-as randomly assigned. Columns differ by the teacher characteristic used for T_{ig}, and rows differ by the baseline student characteristic used for Z_{ig}. Standard errors, reported in parentheses, are two-way clustered at the course-period-year (the strata within which "treatment" is assigned) and student (to account for the fact that students may appear in more than one strata) levels. The statistic at the bottom reports the p-value from overall joint test of balance across all 64 coefficients. A visual representation of this table is given in figure 1. Significance stars denote $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A.2: Measurement-error adjustment and sensitivity of classroom-effect slopes

	Bachelor's attainment	**Log wage at 25**
Panel A. Raw regression summary		
N	1,057	453
R^2	0.034	0.024
$\hat{\beta}$	0.089	0.163
$SE(\hat{\beta})$	0.014	0.047
Panel B. X-axis measurement-error adjustment		
$\hat{\sigma}_x^2$	0.022	0.034
$\overline{\widehat{SE}_x^2}$	0.038	0.039
$\hat{\lambda}_x = \hat{\sigma}_x^2/(\hat{\sigma}_x^2 + \overline{\widehat{SE}_x^2})$	$0.022/(0.022+0.038) = 0.370$	$0.034/(0.034+0.039) = 0.465$
$1/\hat{\lambda}_x$	$1/0.370 = 2.703$	$1/0.465 = 2.149$
$\tilde{\beta} = \hat{\beta}/\hat{\lambda}_x$	$0.089/0.370 = 0.242$	$0.163/0.465 = 0.350$
$SE(\tilde{\beta})$	0.037	0.106
Panel C. Correlated-noise sensitivity		
$\bar{m} = C^{-1}\sum_c \widehat{SE}_{x,c}\widehat{SE}_{y,c}$	0.024	0.051
ρ_u^*	0.156	0.095

Notes: This table shows how the estimated relationship between classroom effects on a long-run outcome and classroom effects on test scores (i.e. the slopes in figure 4) changes once the test-score regressor is treated as noisy. In both columns, the x-axis variable is the estimated classroom effect on test scores and the y-axis variable is the indicated outcome. The row for $\hat{\beta}$ reports the raw OLS slope from regressing the y-axis classroom effect on the x-axis classroom effect, with a heteroskedasticity-robust standard error. Panel B then imposes the benchmark assumption that measurement error in the test-score effect is classical and uncorrelated with measurement error in the y-axis outcome. Under that assumption, the x-axis signal variance is $\hat{\sigma}_x^2$, the average estimated x-axis noise variance is $\overline{\widehat{SE}_x^2} = C^{-1}\sum_c \widehat{SE}_{x,c}^2$, the reliability ratio is $\hat{\lambda}_x = \hat{\sigma}_x^2/(\hat{\sigma}_x^2 + \overline{\widehat{SE}_x^2})$, and the attenuation-corrected slope is $\tilde{\beta} = \hat{\beta}/\hat{\lambda}_x$. Panel C then asks how sensitive that corrected slope is to non-classical measurement error by allowing cross-outcome noise correlation. The row for ρ_u^* reports the smallest positive value of ρ_u in $\mathrm{Cov}(u_{x,c}, u_{y,c}) = \rho_u \widehat{SE}_{x,c}\widehat{SE}_{y,c}$ at which $\tilde{\beta}$ first loses 5 percent significance. Standard errors for $\tilde{\beta}$ come from a nonparametric bootstrap over classroom cells while holding the estimated x-axis signal variance fixed.

Table A.3: Benchmark-rate breakdown of the teacher-quality PDV calculation

Teacher load	$\Delta \log w$	**Annual gain / student**	**Students / teacher-year**	**PDV (growth = inflation)**	**PDV (no wage growth)**
Panel A. Direct 1-SD improvement in log-wage classroom effects					
5 x 20	0.082	\$2,429	100	\$7.00M	\$4.73M
6 x 25	0.082	\$2,429	150	\$10.50M	\$7.09M
6 x 30	0.082	\$2,429	180	\$12.60M	\$8.51M
Panel B. 1-SD improvement in test-score classroom effects mapped into log wages					
5 x 20	0.042	\$1,222	100	\$3.52M	\$2.38M
6 x 25	0.042	\$1,222	150	\$5.28M	\$3.57M
6 x 30	0.042	\$1,222	180	\$6.34M	\$4.28M

Notes: This appendix table reports the benchmark calculations behind the teacher-quality PDV exercise. In Panel A, classroom quality is defined as a one-standard-deviation improvement in the distribution of classroom effects on log wages. In Panel B, classroom quality is defined as a one-standard-deviation improvement in the distribution of classroom effects on test scores and then mapped into log wages using the classical measurement-error-adjusted relationship between classroom effects on log wages and classroom effects on test scores. In both panels, the implied annual earnings gain per student is computed as $\bar{w}_{25}(e^{\Delta \log w} - 1)$, where $\bar{w}_{25}$ is the CyFair mean winsorized wage at age 25. Total PDV for one teacher-year is then $N_T \times \bar{w}_{25}(e^{\Delta \log w} - 1) \times \sum_{a=1}^{40} \psi_a (1+r)^{-a}$, where N_T is students taught per teacher-year, ψ_a is the survival probability at age 25 + a, and r is the annual real discount rate. The benchmark net discount rate is 1.26% and is computed using the monthly average of the 10-year US Treasury yield net of the expected inflation rate for the April 1996 to March 2026 period. The discount rate for PDV calculation under a no wage growth scenario is 3.65% and is computed using the monthly average of the 10-year US Treasury yield for the April 1996 to March 2026 period. The survival probability comes from the Social Security Administration actuarial life tables. 10-year US Treasury yield are obtained from the Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity series in FRED. The expected inflation is obtained from the 26-Year Expected Inflation series in FRED. The small, medium, and large scenarios correspond to 5 classes of 20 students, 6 classes of 25 students, and 6 classes of 30 students, respectively.