ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 24-965

Teachers in our Midst: Using Experienced School Staff to Solve Teacher Shortages

Mary E. Laski
Harvard University

Teacher shortages are a persistent challenge in the United States. I evaluate the effectiveness of an innovative pilot program that allowed principals to hand-select experienced staff members and paraeducators already working in schools to lead classrooms. Pilot educators are predominantly Black or African American. Districts reported randomly assigning students to teachers, and my analysis cannot reject randomization. Controlling for demographics and baseline scores, I find that students assigned to these pilot teachers perform just as well as those assigned to traditionally licensed teachers on average and outperform their peers in math. My results point to an untapped resource of potential teachers and underscore the value of principals' local knowledge to identify capable candidates for teaching positions.

VERSION: May 2024

Suggested citation: Laski, Mary E.. (2024). Teachers in our Midst: Using Experienced School Staff to Solve Teacher Shortages. (EdWorkingPaper: 24-965). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/n1hz-f967

Teachers in our Midst:
Using Experienced School Staff to Solve Teacher Shortages

Mary E. Laski[1]

Last update: May 3, 2024

Abstract

Teacher shortages are a persistent challenge in the United States. I evaluate the effectiveness of an innovative pilot program that allowed principals to hand-select experienced staff members and paraeducators already working in schools to lead classrooms. Pilot educators are predominantly Black or African American. Districts reported randomly assigning students to teachers, and my analysis cannot reject randomization. Controlling for demographics and baseline scores, I find that students assigned to these pilot teachers perform just as well as those assigned to traditionally licensed teachers on average and outperform their peers in math. My results point to an untapped resource of potential teachers and underscore the value of principals' local knowledge to identify capable candidates for teaching positions.

[1] *Mary Laski is Research Principal at the Center on Reinventing Public Education at Arizona State University and a Ph.D. candidate at the Harvard Graduate School of Education (mary.laski@asu.edu).*

I thank the Mississippi Department of Education for its collaboration and support throughout this research. Virginia Lovison was critical in setting up this partnership and advising on the work. Chris Avery, Tom Kane, Dick Murnane, and Eric Taylor all provided invaluable feedback. I also thank participants at the HGSE Doctoral Colloquium and the AEFP, AERA, APPAM, and CESifo Junior Workshop on the Economics of Education conferences for helpful comments and feedback. The opinions expressed are those of the author and do not represent the views of the Mississippi Department of Education. All errors are my own.

Adequately staffing classrooms with effective educators is notoriously difficult, particularly in rural regions, low-income schools, and high-needs subjects (Espinoza et al., 2018; Nguyen et al., 2022). Public school classrooms often include valuable staff members that are *not* the official teacher of record – paraprofessionals, instructional aides, teaching assistants, and so on (Bisht et al., 2021). This paraeducator workforce is rapidly expanding in the US, and offers a large, untapped recruiting pool of potential teachers. This paper examines the impact of allowing principals in high-needs districts to bump up experienced school staff members to full teachers of record to mitigate teacher shortages. Districts report randomly assigning students to teachers, and my analysis cannot reject randomization. I find that, in this context, these educators perform just as well as comparable, traditionally licensed teachers on a wide variety of outcomes, and outperform their peers in math. These educators are also more likely than their peers to stay in the profession in subsequent years and have similar racial demographics to their students. My findings suggest that this program could be an innovative solution to the widespread problem of teacher shortages, particularly in high-needs districts.

I study a three-year pilot of an alternative performance-based licensure (PBL) program introduced in Mississippi in 2019. The PBL program enabled principals to identify experienced staff members or paraeducators in their school that did not pass traditional licensing exams and promote them to a regular, full-time teacher role on a provisional license. Candidates were required to hold a Bachelor's degree and fulfill every other prerequisite to become a teacher except for passing the licensure test. This program is unique in that it is not a universal waiver of traditional licensing requirements, but rather an opportunity for principals to use their own professional judgment to hand-select and reward talented staff who are already working in their

schools. In brief, the program offered experienced school staff a pathway into teaching without requiring traditional licensure testing.

Eight school districts experiencing acute workforce challenges participated in the three-year pilot. In total, 126 educators across sixty-three schools were promoted to full-time teaching roles. All but one educator selected for the PBL pilot program identified as Black or African American, and the median PBL candidate reported roughly seven years of experience working in their school. I observe three cohorts of teachers, with 68 teachers selected in Year 1, 31 teachers in Year 2, and 27 teachers in Year 3.

I ask how PBL candidates perform in the classroom relative to other teachers holding similar teaching positions, including traditionally licensed teachers. As a condition for participating in the pilot program, principals were required to hand-select a comparison teacher for each PBL candidate and randomly assign students across these classrooms. This was not a significant departure from business-as-usual student scheduling, as even in the absence of the PBL program, districts in the study sample use a software program to randomly assign students to classrooms. In practice, students sometimes switch classrooms after initial random assignment, so I stop short of describing the empirical setting as a randomized control design, but I cannot reject that the final student assignments are random. As a conservative check of my estimates, I construct three different counterfactual groups and examine how PBL candidates performed relative to all three. My key outcome measures are student test scores and absences; I also provide some suggestive evidence on teacher observation scores, retention rates, and survey responses.

Across the board, I find that PBL candidates are at least as effective as their peer teachers, on average. PBL candidates also outperform their peers in some areas. Students of PBL

candidates score roughly 0.2 standard deviations higher on math tests than their peers in the same grade and subject in their school, and this difference is statistically significant (p=0.004). Though this estimate comes from a relatively small subsample of PBL teachers (25 treated teacher-years), it represents a staggeringly large effect size: roughly twice the expected impact of being assigned an experienced teacher rather than a novice (Staiger & Rockoff, 2010). Students of PBL candidates are also absent roughly 1 day less than students of hand-selected comparison teachers (p=0.008). On classroom observations, PBL candidates score roughly 0.12 points higher (about a quarter of a standard deviation) than emergency-licensed teachers in the same schools (p=0.016), and these differences are generally consistent across all observed standards. Finally, PBL candidates are significantly more likely than their peers to still be teaching in a PBL district in subsequent years.

This research contributes novel and timely evidence on how experienced school staff perform when promoted to a full-time teaching role. Despite the rapid expansion of the paraeducator workforce in the United States, relatively little is known about these educators (Bisht et al., 2021; Theobald et al., 2023). Encouraging research from North Carolina suggests that paraeducators improve student outcomes (Hemelt et al., 2021), but these effects speak to the combined effect of paraeducators working in conjunction with a traditionally licensed teacher, an important distinction from the PBL model.

I also contribute to a large body of work on the value of traditional teacher licensing exams. A common reason individuals work in paraeducator or unlicensed educator roles rather than regular teaching positions is because they struggle to obtain a teaching license through traditional pathways. Nearly every US state requires teacher candidates to pass a licensure exam. Proponents of traditional teacher licensure exams argue these exams help uphold minimum

standards for the teaching profession. However, these exams may disproportionately screen out prospective educators of color from the teaching workforce (Nettles et al., 2011). These licensure exams also privilege academic skills over relational skills, even though both are central to student success. Goldhaber & Hansen (2010) examine the relationship between teacher licensure scores, teacher identity, and student performance and find that Black students learn just as much with lower-scoring Black teachers as they do with higher-scoring White teachers. These results suggest that licensure tests may be a particularly poor signal of Black teacher effectiveness, or that any relationship between test scores and effectiveness could be offset by teacher-student race match effects.

Finally, this paper also contributes to a rich body of work seeking to understand and explore opportunities to mitigate teacher shortages (Cowan et al., 2016; Dee & Goldhaber, 2017; García & Weiss, 2020; Goldhaber et al., 2015; Ingersoll et al., 2019; Podolsky et al., 2016). In particular, the PBL program is relevant to the large literature on alternative certification programs such as Teach for America (TFA). Previous research finds that TFA teachers are more effective than traditionally licensed teachers, but also have much higher turnover rates (Kane et al., 2008; Lovison, 2022). The PBL program differs from the typical alternative teacher licensure programs which focus on recruiting talented individuals with no background in education to serve short-term stints in schools. Rather, the focus here is on experienced educators who have already worked in the schools where they will teach but have previously been excluded from teaching positions based purely on credentialing issues. Unlike typical alternatively licensed teachers, PBL candidates have the potential to be relatively effective due to their extensive experience in schools *and* be less likely to leave teaching due to their pre-existing connections with the community. My analyses confirm these hypotheses.

This paper is organized as follows. Section I describes the setting for the study and the available data. Section II outlines my hypotheses and a simple framework describing potential mechanisms for program effects. Section III describes my empirical approach, and Section IV details the empirical results. In Section V, I conclude by discussing the implications of these results.

I. Setting and Data

This program was piloted in Mississippi, a state in the American South with a population of 2.95 million people and nearly 450,000 students enrolled in public schools. Students in Mississippi public schools are predominantly low-income: nearly 75% are eligible for free or reduced-price lunch. The Mississippi Department of Education (MDE) designed this pilot program with the goals of supporting high-needs districts in filling vacancies and increasing the diversity of the teacher workforce. MDE selected four districts to participate in the pilot based on existing state priorities, and four more districts were recruited through a formal application process. Eligible districts were required to attend a focus group meeting about the program, submit a formal application, and consent to randomly assign students across PBL candidates and comparison teachers. Seventy-three districts indicated interest in the pilot; four additional districts were ultimately selected for participation based on needs and capacity.

Participating districts are generally high-needs: students are majority low-income and schools are particularly hard to staff. Figure 1 presents district-level responses to a state survey on teacher vacancies. There is considerable variation in vacancy rates across the state, but pilot districts (shaded pink bars) generally had higher rates of teacher departures in the 2020-21 school year (top panel) and higher vacancy rates in fall 2021 (bottom panel). According to state report

card data, in 2020-21, pilot districts had anywhere from 15% to 49% of positions covered by emergency or provisional licenses (Mississippi Succeeds Report Card, 2021).

Principals in pilot districts were invited to select promising educators already working in their schools to participate. PBL candidates could bypass the PRAXIS licensure requirement and assume a position as a full-time teacher with a provisional license for three pilot years, with the promise of a full license if they demonstrated effectiveness during the pilot. Principals hand-selected promising educators with a track record of success fulfilling other teaching-focused roles to be in the pilot program. PBL candidates previously worked as paraeducators, teacher assistants, instructional aides, and similar roles. All but one PBL candidate identified as Black or African American. Principals also hand-selected a comparison teacher in their school for each PBL teacher. Comparison teachers had to be traditionally licensed and, whenever possible, taught in the same subject and grade as the PBL candidate.[2] Participating districts and principals agreed to randomly assign students to PBL candidates and comparison teachers, though this was only possible when the comparison teacher was teaching in the same grade and subject. PBL candidates were assigned teaching slots spanning Pre-K through high school. About one-fifth of PBL candidates led Pre-K or Kindergarten classrooms. Roughly one-half of PBL candidates worked in grades 3 or higher.

To study the effectiveness of the PBL pilot, I use detailed administrative data on students and teachers provided by MDE for school years 2015-16 to 2021-22. Student-level data includes

[2] District PBL Coordinators were asked to work with principals in selecting comparison teachers according to ordered criteria: (1) standard licensure, (2) same/similar grade level, (3) same/similar content area, (4) same/similar years of experience. Comparison teachers *had* to meet the first item (standard licensure) and they should then meet as much of the criteria from there, with an understanding that in some high-need areas the first item alone might mean finding a comparison teacher in a different grade (since all other teachers had provisional licenses). About 55% of pilot-comparison pairs actually taught in the same grade and subject. 80% taught in the same subject(s); 63% taught in the same grade.

scores on annual standardized tests, demographics, absences, and course schedules linking students to teachers. Teacher-level data includes demographic and licensure information. I also have annual average teacher observation scores for all teachers in pilot districts in the 2021-22 school year. Finally, I also have data from two surveys run by the state department: a survey of all districts gathering information on teacher vacancies, and annual surveys of PBL candidates and their hand-selected comparison teachers gathering information on the program and their future career plans.

Table 1 presents basic descriptive statistics on students and teachers in the treated PBL pilot group (column 4) and various comparison groups (columns 1 through 3; see Section III for details on comparison groups). I present data on two samples: all students and educators where I have attendance data (top panel) and the smaller subset of students and educators where I have test score data (bottom panel). In all groups, students and teachers are majority Black, though 100% of PBL candidates in the first-year sample identify as Black.[3] Teachers in all groups are majority female. PBL candidates have much less experience than their hand-selected comparison teachers and teachers in the same school, grade, and subject, but generally have more years of experience than emergency licensed teachers in the same district. I note that the years of experience variables are inconsistently documented and do not always follow expected trajectories, so I caution over-interpreting them. I do not include these variables in my main analyses given my concerns about the data quality.

[3] In the full sample, 125 of 126 PBL candidates (99.2%) identify as Black or African American. Table 1 presents pre-treatment data from 2020; not all PBL candidates were working in schools at this time, which is why the numbers differ slightly.

II. Hypotheses & Theoretical Framework

Teachers' effectiveness at improving student outcomes varies widely (Kane et al., 2008; Koedel et al., 2015). However, the specific determinants of teacher performance are still unclear, and few observable characteristics reliably predict teacher effectiveness (Rockoff et al., 2011; Staiger & Rockoff, 2010). Specifically, teacher certification status does not appear to be meaningfully related to student achievement (Kane et al., 2008). Here, I review various additional drivers of teacher effectiveness and how these might relate to the study of this pilot program. I conclude that the evidence-based prediction of PBL candidates' relative performance is ambiguous, given the combination of these opposing mechanisms.

While graduate degrees and certifications rarely predict meaningful differences in teacher effectiveness (Clotfelter et al., 2010; Goldhaber & Brewer, 1997; Staiger & Rockoff, 2010; Rivkin et al., 2005), there are several studies that utilize more detailed data on previous academic performance and cognitive ability and find strong predictive power on effectiveness (Rockoff et al., 2011; Jacob et al., 2018; Taylor, 2018). If the PRAXIS exam required of the traditional licensure pathway truly distinguished potential candidates based on some measure of cognitive ability, then PBL candidates are likely *lower*-performing than their traditionally licensed peers, as they did not pass this cognitive benchmark.

Teacher experience is the most reliable and consistent predictor of teacher effectiveness (Rockoff, 2004; Papay & Kraft, 2015). PBL candidates generally have fewer documented years of teaching experience than their traditionally licensed peers (see Table 1), suggesting they would be *lower*-performing. However, I note two caveats to this simple interpretation. First, both PBL candidates and comparison teachers have over five years of documented experience, on average, and previous research suggests that returns to experience are particularly steep in the

first one to five years of experience and then flatten out later in the career (Papay & Kraft, 2015). As such, the differences in experience documented here may not be particularly meaningful. Second, to be eligible for this program, PBL candidates had extensive experience in their specific schools – not leading classrooms, but regularly assisting and aiding teachers, and often performing roles as longer-term substitutes. Most existing estimates of returns to experience combine learning about the school community and learning how to lead one's own classroom, which are inherently difficult to disentangle. I argue that PBL candidates' more local knowledge and experience with their specific student body could be particularly valuable, and thus differences in experience lead to an *ambiguous* prediction of the relative effectiveness of PBL candidates.

This local experience connects to some level of non-academic or relational ability that is likely not captured via traditional licensure pathways but could potentially be understood by a principal with knowledge of their own staff. This local knowledge mechanism is understudied compared to the other mechanisms discussed here. In a context of asymmetric and imperfect information, principals may hold particular knowledge on local staff members' potential for effectiveness, and thus can correctly identify and recruit the highest-potential educators for the pilot program. This principal information mechanism would suggest PBL candidates would be *at least as effective as traditionally licensed teachers, if not higher-performing*.

Finally, a large body of evidence demonstrates that students of color benefit academically from having a teacher of the same race (Dee, 2004, 2005; Egalite et al., 2015; Harbatkin, 2021). Same-race teachers also tend to have higher expectations and more positive beliefs about their students (Gershenson et al., 2016; Dee, 2005). More recent evidence demonstrates that, for Black students, being randomly assigned to a Black teacher in grades K–3 significantly increases the

probability of graduating high school and enrolling in college (Gershenson et al., 2019). Over 95% of students in participating districts and 99% of PBL candidates identify as Black or African American. Roughly 80% of non-PBL teachers in participating districts identify as Black or African American. Given the extensive evidence on positive impacts of student-teacher race match, these differences suggest PBL candidates would be *higher*-performing than their average traditionally licensed peer.

Thus, I conceptualize teaching effectiveness as related to a bundle of observable and unobservable traits, including cognitive ability, experience (generally and locally), non-cognitive or relational ability, and racial congruence with students. Given this bundle of traits, the hypothesized relative effectiveness of PBL candidates is ambiguous: even if the PBL teachers are weaker in terms of certification test scores, they may be stronger in terms of their match with the students in the schools and their knowledge of the specific school environment.

III. Empirical Approach

To understand how teachers participating in the PBL pilot performed relative to their counterparts, I use three different comparison groups:

1) Hand-selected group. At the outset of the project, program staff hand-selected a suitable comparison teacher for every PBL candidate. Hand-selected comparison teachers were fully licensed and taught a tested grade or subject in the same school as the PBL candidate. Whenever possible, these teachers taught in the same grade and subject as the PBL candidate. The program was designed such that classroom rosters were randomly assigned to PBL candidates and their comparison teacher, though this was only possible when the comparison teacher taught in the same grade and subject (roughly 55% of pairs).

2) The randomly assigned group. Classroom assignments in pilot districts are randomly generated within school, grade, and subject by scheduling software. Principals can amend assignments after the initial randomization, but this is allegedly uncommon, according to several district superintendents. Balance tests suggest any observable differences between the classrooms of PBL candidates and their colleagues in the same school, grade, and subject are minimal, and I cannot reject that classroom assignments are random.
3) The policy-relevant counterfactual group. Conversations with superintendents revealed the hand-selected comparison teachers are generally strong performers, and as such they are potentially poor proxies for the teachers that the districts *would have been able to hire in the absence of the pilot program*. For a more policy-relevant analysis, I compare the performance of teachers selected for the pilot program to teachers working on emergency or provisional licenses in similar schools – according to superintendents, if not for this PBL pilot program, these vacancies would likely all be filled by educators with emergency and/or out-of-field certifications.

To estimate effects on student outcomes, I fit models of the following form:

$$A_{ijt} = \beta Pilot_j + \theta A_{it-1} + \delta \mathbf{X}_i + \pi_p + \tau_c + \varepsilon_{ijt} \quad (1a)$$

$$A_{ijt} = \beta Pilot_j + \theta A_{it-1} + \delta \mathbf{X}_i + \pi_{sgz} + \varepsilon_{ijt} \quad (1b)$$

$$A_{ijt} = \beta Pilot_j + \theta A_{it-1} + \delta \mathbf{X}_i + \pi_s + \tau_g + \varepsilon_{ijt} \quad (1c)$$

where A_{ijt} represents an outcome (achievement or absences) for student *i* with teacher *j* in year *t*. $Pilot_j$ is a teacher-level binary variable equal to 1 for PBL pilot candidates and 0 for the comparison group of interest. To utilize the three different comparison groups outlined above, I include different fixed effects in each model. Models using hand-selected comparison

teachers (1a) include pilot-pair *p* and cohort *c* fixed effects to compare each PBL candidate to their own specific comparison teacher. Models using the potentially randomly-assigned group (1b) use school-by-grade-by-subject *sgz* fixed effects to compare each PBL candidate to teachers in their own grade and subject, where classroom rosters were randomized. Models comparing to emergency licenses (1c) use separate school *s* and grade *g* fixed effects.

All models include controls for both previous-year test scores and previous-year absences.[4] $\mathbf{X}_i$ represents a vector of student-level controls such as gender, racial demographics, Title I eligibility, English language learner status, eligibility for free or reduced-price lunch, and grade fixed effects. Test score analyses also include subject fixed effects. Test scores are standardized within test, grade, and subject using the statewide distribution such that the mean is 0 and the standard deviation is 1. Standard errors are clustered at the teacher level (or the teacher-year level when pooling across years). Given the distinct nature of PBL districts, all samples are limited to treated districts. Test score analyses limit to classes with at least 5 students, though the findings are robust to lifting this restriction.

Given the fact that emergency-licensed teachers are likely the policy counterfactual of interest, I also fit models of the following form:

$$A_{ijt} = \beta Pilot_j + \gamma Emergency_j + \theta A_{it-1} + \delta \mathbf{X}_i + \pi_s + \tau_g + \varepsilon_{ijt} \quad (2)$$

where $Emergency_j$ is a teacher-level binary variable equal to 1 for emergency-licensed teachers in treated districts and 0 otherwise and π_s and τ_g represent separate school and grade fixed effects. By including both $Pilot_j$ and $Emergency_j$ in the same model, I can separately identify

[4] Lagged test scores are the average of math and ELA test scores in the most recent tested year. Mississippi did not conduct traditional standardized tests in 2020 due to the COVID-19 pandemic. For lagged test scores in the 2020-21 data, I use data from 2018-19 whenever available. If lagged test scores or absences are missing, the average value is imputed, and all models include a control for the missing lags.

PBL candidates' and emergency-licensed teachers' differential effects on student outcomes, relative to traditionally licensed teachers in treated districts. In tables presenting my main effect estimates, I present the difference between β and γ from these models as well as the results of an F-test of coefficient equivalency. This test essentially indicates whether PBL candidates' performance is statistically different than emergency-licensed teachers, relative to traditionally licensed teachers in treated districts.

To further study how PBL candidates perform on student achievement and investigate the relative variation in program effects, I estimate a simple teacher-year value-added model with teacher-year-specific random intercepts α_{jt}, as follows:

$$A_{ijt} = \alpha_{jt} + \theta A_{it-1} + \delta \mathbf{X}_i + \tau_g + \theta_z + \vartheta_d + \varepsilon_{ijt} \qquad \text{where } \alpha_{jt} \sim N\left(0, \sigma^2_{\alpha_{jt}}\right) \tag{3}$$

I include fixed effects for grade (τ_g), subject (θ_z), and district (ϑ_d) as well as lagged outcomes and the standard vector of student-level controls outlined above. I plot the predicted values of the teacher-year-specific intercepts ($\widehat{\alpha_{jt}}$) separately by licensure status, illustrating the distribution of teacher-year-specific estimates of value-added.

I also compare annual averages of classroom observation scores across teacher groups. All teachers in Mississippi are subject to at least three observations annually. Observers (usually school principals) rate teachers on nine standards in four domains on a scale of 1 to 4. Previous research on similar observation schemes has demonstrated the predictive validity of these measures; in short, classroom observations are predictive of teacher performance and student achievement (Bacher-Hicks et al., 2019; Kane & Staiger, 2012). Figure 2 presents the detailed rubric with all nine standards. The first and second domains focus specifically on lesson content and student learning, which I argue are the standards most closely aligned with the content

knowledge measured in traditional licensure exams. The third domain covers classroom management and environment, and the fourth domain covers professionalism. In pilot districts, average scores for each standard are roughly 3 (ranging from 2.8 to 3.1), with a standard deviation of roughly 0.6 (ranging from 0.45 to 0.65). I have annual average observation scores for all observed teachers in pilot districts in 2021-22. I fit models of the following form separately for each standard and for the summative average of the nine standards:

$$O_j = \beta Pilot_j + \theta_s + \varepsilon_j \quad (4)$$

where O_j is an observation outcome (individual standard or summative average) for teacher *j*, and θ_s represents school fixed effects.[5] I note that my analyses are limited here as I only have one year of data, and observations are often conducted by principals, who may be motivated to rate PBL candidates particularly highly to support their path to alternative licensure.

Finally, I compare retention rates across teacher groups. I fit models of the following form:

$$Retained_{jt} = \beta Pilot_{j,t-1} + \theta_d + \varepsilon_{jt} \quad (5)$$

where $Retained_{jdt}$ is a binary equal to 1 if teacher *j* is still teaching in any PBL district *d* in year *t*. $Pilot_{j,t-1}$ is a binary equal to 1 for PBL candidates in the previous year and 0 for the comparison group of interest. Analyses are limited to PBL districts and include district fixed effects.

Pre-Treatment Balance

Table 2 presents traditional tests of pre-treatment balance on student covariates across groups. Each cell represents a coefficient from a separate regression with the covariate of interest

[5] Many teachers teach multiple grades, which is why I do not include grade fixed effects in this teacher-level specification.

as the outcome and a binary treatment variable as the predictor of interest, following equation (1). While most baseline covariates are not meaningfully different across groups, as seen in Table 1, I do note a few statistically significant differences in student race across classrooms, despite the differences themselves being quite small (often a fraction of a percentage point). I include controls for these racial demographics in all models. I also separately test for joint orthogonality by fitting models with the binary $Pilot_j$ indicator as the outcome and the entire set of relevant student-level covariates as predictors in the same model. When conducting F-tests of the joint hypothesis that all coefficients equal zero, I consistently estimate F-statistics below 1.9 and p-values above 0.05 for the school-grade-subject comparison group (column 2), meaning I cannot reject the null hypothesis that all coefficients are zero and thus that rosters were randomized within this group.

The main threat to identification in this analysis would be principals amending randomized student assignments based on students' potential for growth on test scores (or improvement in absence rates). If this were to happen, my models would attribute these differences in growth to differences in educators' effectiveness. While potential for growth is inherently unobservable, I argue that the baseline equivalence on previous test scores and previous absences is reassuring: across all comparison groups, I find no evidence that these lagged values of key outcomes are statistically different between treated and comparison teachers. This baseline equivalence provides at least suggestive evidence that there was no overt ability tracking between PBL candidates' and comparable teachers' classrooms.

IV. Results

Overall, I find that PBL candidates are at least as effective as their peer teachers, on average. Nearly all estimated differences are not statistically distinguishable from zero, with a

few notable exceptions. Students of PBL candidates are absent roughly 1 day less than students of hand-selected comparison teachers (p=0.008). In math, students of PBL candidates score roughly 0.2 standard deviations higher than their same school-grade-subject peers (p=0.004 for pooled estimate). On classroom observations, PBL candidates score roughly 0.1 points higher (about a sixth of a standard deviation) than emergency-licensed teachers in the same districts (p=0.016), and these differences are generally consistent across all observed standards. Finally, PBL candidates are significantly more likely to still be teaching in a PBL district in subsequent years than other teachers in their district.

Main Effect Estimates

Table 3 presents estimated treatment effects on absences and test scores separately for 2020-21 and 2021-22 and pooled across both years. Columns (1) through (3) present coefficients from models following equation (1); column (4) presents the difference in coefficients and associated F-test details from equation (2). These analyses suggest that, on average, teachers participating in the pilot alternative-licensure program performed on par with comparable teachers working in similar schools in similar teaching positions: the great majority of effect estimates are very close to zero and are not statistically distinguishable from zero. Note that, due to sample sizes and clustered standard errors, some confidence intervals are somewhat large. In most cases, I cannot rule out that PBL candidates are less effective than comparison teachers, but I can rule out large differences. For example, in the randomly assigned comparison group (column 2), I can rule out test score losses larger than 0.05 standard deviations: roughly the difference between a first-year teacher and a third-year teacher (Papay & Kraft, 2015).

These findings are consistent across all three groups of comparison teachers, both outcomes, and both years where I have outcome data. One notable exception is a significant

positive effect of being assigned a PBL candidate on student attendance: when pooling across years, students of PBL candidates were absent roughly 1 day less than students of hand-selected comparison teachers. This represents a meaningful difference: students in these districts are absent roughly 7 to 9 days a year, on average, so this would be a difference in absences of roughly 12%. This difference is significant at the 1% level (p=0.008).

Figure 3 plots the distribution of teacher-year-level value-added estimates on standardized test scores using the random effects model detailed in equation (3). I again find convincing evidence that the distribution of PBL candidates' value-added (red line) is roughly on par with both their hand-selected comparison teachers (dotted line) and the emergency-licensed teachers (solid black line) in pilot districts. Indeed, the mean value-added estimate for PBL candidates is slightly higher than the mean for comparable teacher groups, though this difference is not statistically significant at traditional levels. Note that this plot uses empirical best linear unbiased predictions (EBLUPs), or "shrunken" estimates, so the figure is understating the true variance of the distributions.

I also estimate treatment effects on student achievement separately by test subject. Estimates from these models are presented in Table 4. Here, I find that the overall null effects mask differential effects across subjects. Treated teachers consistently outperform their peers in the same grade and subject in math and generally perform on par with comparable teachers in ELA, on average. In math, students of PBL candidates score roughly 0.2 standard deviations higher on math tests than peers in the same grade, and this difference is statistically significant (p=0.004 for the pooled estimate). Figure 4 plots the distribution of math value-added, again following equation (3), and demonstrates that PBL candidates' value-added estimates is higher, on average, than comparison teachers. Note that this represents a small subsample of the treated

group, as not all PBL candidates taught math: the pooled estimate includes 25 treated teacher-years.[6]

Differences in Teacher Outcomes

I also document meaningful differences in teacher outcomes between PBL candidates and comparable teachers. I begin with differences in observation scores, documented in Table 5. Like the tables above, columns (1) through (3) present coefficients from models following equation (4). Column (4) presents the difference between coefficients on treatment and emergency license indicators and the associated F-test details. I look at differences in the summative evaluation score for the year (row 1) as well as the annual average score for each individual standard (rows 2 through 10). See Figure 2 for the detailed observation rubric.

I find that, on average, PBL candidates score about the same as their hand-selected comparison teachers and the teachers in their same school. Indeed, the summative scores for PBL candidates are generally roughly 0.07 points higher than scores for teachers in the same school, and that difference is significant at the 10% level (p=0.056). These differences seem to be driven by relatively large differences in classroom management (standard 5) and engagement in professional learning (standard 8). When comparing PBL candidates to emergency-licensed teachers in the same school, I consistently see large, statistically significant differences. PBL candidates outscore these comparable teachers on nearly every standard by roughly 0.12 points, or about one quarter of a standard deviation, and most of these differences are statistically significant.

[6] This includes 17 unique PBL teachers, roughly 15% of the treated sample.

The fact that these effects are relatively consistent across standards is particularly compelling: PBL candidates are not simply better at classroom management than their peers (standards 5-7), but also perform on par with or outperform their peers in standards that are explicitly focused on lesson content and student learning (standards 1-4). These content-heavy standards are theoretically most closely aligned with the content knowledge measured through traditional licensure exams. While the PBL program allowed these candidates to bypass these testing requirements, it is clear they are not being outperformed by traditionally licensed teachers in content-specific areas, at least based on principals' perceptions of their performance. I note again that observations are often conducted by principals, who may be uniquely motivated to keep PBL candidates in the classroom and thus rate them highly. Still, I find these differences reassuring and aligned with the findings that PBL candidates are generally at least as effective as comparable teachers on improving student outcomes.

I next turn to teacher retention rates. Effect estimates are presented in Table 6. First, I find that PBL candidates are just as likely to still be teaching in a PBL district as their hand-selected comparison teachers: the difference in retention rates between these two groups is quite small and not statistically significant (column 1). When looking at all teachers in PBL districts (column 2), I find that PBL candidates are a full 10 percentage points more likely to still be teaching in both the 2021-22 and 2022-23 school year than traditionally licensed teachers ($p=0.017$ for 2021-22; $p=0.007$ for 2022-23). This is a large, meaningful difference, as the retention rates among traditionally licensed teachers in these districts is roughly 80%. I also find that PBL candidates are about 8 percentage points more likely than emergency-licensed teachers to remain teaching in 2021-22 ($p=0.090$) and 16 percentage points more likely to still be teaching in 2022-23 ($p<0.001$).

Finally, Figure 5 presents responses of PBL candidates and hand-selected comparison teachers on a program survey on longer-term career interests. I find that the great majority of PBL candidates – over 80% – agree to some extent that they view teaching as their long-term career. This is about the same rate of agreement as hand-selected comparison teachers, though PBL candidates are somewhat more likely to strongly agree. I also find that PBL candidates are about as likely as comparison teachers to agree to some extent that they can see themselves teaching in their district for the rest of their career (67% comparison, 60% PBL candidates).[7] These survey responses suggest that, if the short-term effect estimates outlined above can persist, this alternative licensure program could potentially provide a valuable long-term solution to the serious problem of teacher shortages.

Sensitivity Analyses

The above estimates could be biased if students are non-randomly assigned to teachers. While the sample districts use a software program to randomly assign students to classrooms, post-randomization amendments are possible, and I lack detailed data on the initial randomization. I also found some small but significant differences in observable characteristics across classrooms in Table 2, which could be cause for concern. In my test score analyses, non-random student assignment to teachers would be particularly problematic for my interpretation if principals assigned PBL candidates the students most likely to improve their test scores. In this scenario, I would attribute students' achievement growth as a PBL effect. I note that I do not find any significant differences in students' previous test scores, so sorting of this type seems unlikely, but is still theoretically possible.

[7] Note that neither of these reported differences are statistically significant by traditional standards ($p>0.05$).

To address this potential bias, I can look broadly at any grade level where a PBL candidate teaches. This analysis alleviates sorting concerns by identifying all students in a grade as treated, not just the students assigned to a PBL candidate's classroom. To do this, I fit a difference-in-difference model at the grade-level using two-way fixed effects. A school-grade cell is considered "treated" if there is a PBL candidate teaching in that school in that grade in that year. In brief, I compare grades "treated" with a PBL candidate to all other grades without a PBL candidate, using models of the following form:

$$A_{igst} = \delta Pilot_{gst} + \theta \mathbf{X}_i + \pi_{gs} + \omega_{(g)t} + \varepsilon_{igst} \quad (6)$$

where A_{igst} represents the test score for student *i* in grade *g* in school *s* in year *t*. $Pilot_{gst}$ is a variable indicating whether a school-grade-year cell is "treated" with a PBL candidate. I fit models where $Pilot_{gst}$ is a binary equal to 1 in school-grade-year cells that had a PBL candidate and 0 otherwise. I also fit models where $Pilot_{gst}$ represents the proportion of students in each school-grade-year cell taught by a PBL candidate (for example, if a grade had two equal-sized classes and one was taught by a PBL candidate, $Pilot_{gst}$ would be 0.5 for all observations in that grade). I include the vector of student-level demographics outlined above. All models also include fixed effects for school-grade and for year, and I also fit models that use fixed effects for grade-year. I again limit to treated districts. For these estimates to be interpreted as causal, the identifying assumption is that the difference in outcomes between grades that did and did not have PBL candidates would have remained constant over time in the absence of PBL. Figure 6 presents suggestive evidence that trends in test scores were roughly parallel pre-treatment, though treated grades may have been on a slightly upward trajectory relative to comparison

grades before treatment (and the onset of the pandemic, which notably decreases scores post-treatment).

Table 7 presents results from this line of analysis. I find no evidence that treated grades are significantly different than untreated grades after implementation of PBL. This is generally in line with my main finding on pooled test results, though lacks precision. When looking specifically at math, I find generally positive, marginally significant estimates that are much smaller than my preferred estimates. That said, the confidence intervals prevent us from rejecting effect estimates on math scores as large as 0.054 standard deviations when using the binary treatment indicator. While this estimation is generally less precise than my preferred specification, the fact that estimates are trending in similar directions suggests that the main findings are not notably biased by student sorting across classrooms.

V. Discussion and Conclusion

I study a novel performance-based licensure (PBL) program, which utilizes a previously untapped resource to address the persistent problem of teacher shortages. I find that, on average, PBL candidates perform just as well as various comparable groups of teachers, with a few notable exceptions. PBL candidates outperform teachers in their same grade and subject in math; students of PBL candidates score roughly 0.2 standard deviations higher on math tests than their peers in the same grade and subject – a large and consistently significant effect. Students of PBL candidates also seem to be absent somewhat less than students of hand-selected comparison teachers, and PBL candidates score roughly 0.12 points higher (about a quarter of a standard deviation) on classroom observations than emergency-licensed teachers in the same districts. PBL candidates are also significantly more likely than comparable teachers in their district to remain teaching in the district in subsequent years.

Revisiting the hypotheses outlined in Section II, I find these to be both reasonable and reassuring outcomes for this pilot program. It was possible that PBL candidates would ultimately be lower-performing than the average Mississippi educator, as they are not traditionally licensed and are relatively new to leading instruction. However, I see no evidence of this: I find very few negative effects, and the ones I do find are small and rarely statistically significant. Rather, my findings suggest that the traditional licensing pathway may be arbitrarily preventing capable teachers of color from entering the profession. This suggestion is further supported by the findings on teacher observation scores, as PBL candidates do not score lower on the standards most closely aligned with the content knowledge usually measured on traditional licensure tests (or on any of the observed standards).

To put these effects into context, it's helpful to consider them in comparison to other traditional strategies for filling teacher vacancies. Vacancies are ideally filled by traditionally licensed novice teachers, though it is well-established that these teachers are much less effective than their more experienced peers (Papay & Kraft, 2015; Staiger & Rockoff, 2010). Other alternative teacher licensure programs such as Teach for America (TFA) focus on recruiting talented individuals with no background in education to serve short-term stints in schools. While TFA teachers are more effective initially than traditionally licensed novices, these benefits are somewhat offset by TFA teachers' substantially higher turnover rates (Kane et al., 2008; Lovison, 2022). Evidence on efforts to increase educator diversity also finds that teachers of color have significantly higher turnover rates than white teachers, often due to poor working conditions at their schools (Ingersoll et al., 2019). In stark contrast to these alternatives, I find that PBL teachers are just as effective as traditionally licensed experienced teachers on average, are just as likely (if not more likely) to remain teaching in the district in subsequent years, *and*

are much more likely to be the same race/ethnicity as their students. The unique combination of relative effectiveness, retention, and racial diversity underscores the unique value of the PBL program in filling teacher vacancies, especially in high-needs districts.

This paper provides convincing evidence that PBL candidates are just as effective, if not more effective, than comparable teachers in their district. When considering the potentials of expanding similar programs, it is important to clarify the limitations of the current study and why these effects may not scale. One major limitation is that this program was piloted in eight high-needs districts, and there may be a particular oversupply of potential candidates in these specific contexts. There is not currently any evidence that a program like this would work in a different type of district. Another potential concern is that PBL candidates are the "low-hanging fruit:" the current candidates are a particularly effective stock of potential teachers, but there is not a sufficienct inflow of future PBL candidates. I provide some suggestive evidence that this is not the case given the three cohorts studied here, but supply may decline further in future years. Another concern is what PBL candidates would have been doing if not for PBL. These candidates were already working in their school in another role, and thus PBL creates a new staff vacancy while filling a teacher vacancy. Evidence from other states suggests increasing paraeducator attrition in recent years (Theobald et al., 2023), underscoring the need for careful consideration about the potential costs involved in moving candidates from paraeducator to teacher of record. Finally, expansion of the program could eventually change selection into teaching or pre-PBL jobs. The direction of this selection is theoretically unclear but could potentially change the implications of the program. In short, further research is needed to assess how PBL-like programs could potentially reduce shortages in broader terms.

Overall, my results suggest that the licensure exams in Mississippi may be arbitrarily limiting the supply of effective teachers, and particularly teachers of color. On average, the PBL candidates I studied were just as effective as other teachers working in similar roles, and indeed outperformed similar teachers in some areas. At minimum, the evidence suggests that principals' professional judgment may be a suitable replacement for traditional licensing exams for members of the school community who want to become full-time teachers, but do not yet have the necessary credentials. The suggestive positive effects I find are encouraging, and in line with the premise that unlike an exam, principals can identify teachers with the relational capacity to make a positive impact on students.

References

Bacher-Hicks, A., Chin, M., Kane, T., & Staiger, D. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, *73*, 101919.

Bisht, B., LeClair, Z., Loeb, S., & Sun, M. (2021). Paraeducators: Growth, Diversity and a Dearth of Professional Supports. *Annenberg Institute at Brown University: EdWorkingPaper 21*-490.

Cowan, J., Goldhaber, D., Hayes, K., & Theobald, R. (2016). Missing elements in the discussion of teacher shortages. *Educational Researcher, 45*(8), 460–462.

Clotfelter, C., Ladd, H., & Vigdor, J. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources, 45*(3), 655-681.

Dee, T. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics, 86*(1), 195–210.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review, 95*(2), 158–165.

Dee, T., & Goldhaber, D. (2017). Understanding and addressing teacher shortages in the United States. *The Brookings Institution.*

Egalite, A., Kisida, B., & Winters, M. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review, 45*, 44–52.

Espinoza, D., Saunders, R., Kini, T., & Darling-Hammond, L. (2018). Taking the Long View: State Efforts to Solve Teacher Shortages by Strengthening the Profession. *Learning Policy Institute*.

García, E., & Weiss, E. (2020). Examining the factors that play a role in the teacher shortage crisis (No. 177726). *Economic Policy Institute.*

Gershenson, S., Holt, S., & Papageorge, N. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review, 52*, 209–224.

Gershenson, S., Hart, C., Hyman, J., Lindsay, C., & Papageorge, N. (2019). The long-run impacts of same-race teachers. *Annenberg Institute at Brown University: EdWorkingPaper 19*-43.

Goldhaber, D., & Brewer, D. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, *32*(3), 505-523.

Goldhaber, D., & Hansen, M. (2010). Race, Gender, and Teacher Testing: How Informative a Tool Is Teacher Licensure Testing? *American Educational Research Journal, 47*(1), 218–251.

Goldhaber, D., Krieg, J., Theobald, R., & Brown, N. (2015). Refueling the STEM and special education teacher pipelines. *Phi Delta Kappan, 97*(4), 56–62.

Harbatkin, E. (2021). Does student-teacher race match affect course grades? *Economics of Education Review, 81*.

Hemelt, S. W., Ladd, H. F., & Clifton, C. R. (2021). Do teacher assistants improve student outcomes? Evidence from school funding cutbacks in North Carolina. *Educational Evaluation and Policy Analysis*, *43*(2), 280-304.

Ingersoll, R., May, H., & Collins, G. (2019). Recruitment, employment, retention and the minority teacher shortage. *Education Policy Analysis Archives, 27*(37).

Jacob, B., Rockoff, J., Taylor, E., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics, 166*, 81-97.

Kane, T., Rockoff, J., & Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615–631.

Kane, T., & Staiger, D. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. *MET Project. Bill & Melinda Gates Foundation*.

Koedel, C., Mihaly, K., & Rockoff, J. (2015). Value-added modeling: a review. *Economics of Education Review, 47*, 180–195.

Lovison, V. (2022). The Effects of High-performing, High-turnover Teachers on Long-run Student Achievement: Evidence from Teach For America. *Annenberg Institute at Brown University: EdWorkingPaper 22*-675.

Mississippi Succeeds Report Card. (2021). Available at https://msrc.mdek12.org/.

Nettles, M., Scatton, L., Steinberg, J., & Tyler, L. (2011), Performance and Passing Rate Differences of African American and White Prospective Teachers on PRAXIS Examinations. *ETS Research Report Series*: i-82.

Nguyen, T., Lam, C., & Bruno, P. (2022). Is there a national teacher shortage? A systematic examination of reports of teacher shortages in the United States. *Annenberg Institute at Brown University: EdWorkingPaper 22*-631.

Papay, J., and Kraft, M. (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 (October): 105–19.

Podolsky, A., Tara, K., Bishop, J., & Darling-Hammond, L. (2016). Solving the teacher shortage: How to attract and retain excellent educators. *Learning Policy Institute.*

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review* 94 (2): 247–52.

Rockoff, J., Jacob, B., Kane, T., & Staiger, D. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy, 6*(1), 43-74.

Staiger, D., and Rockoff, J. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives, 24*(3): 97-118.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2019). Understanding teacher shortages: An analysis of teacher supply and demand in the United States. *Education policy analysis archives*, *27*(35).

Taylor, E. (2018). Skills, job tasks, and productivity in teaching. *Journal of Labor Economics, 36*(3), 711-742.

Theobald, R., Kaler, L., Bettini, E., & Jones, N. (2023). A Descriptive Portrait of the Paraeducator Workforce in Washington State. *National Center for Analysis of Longitudinal Data in Education Research (CALDER), Working Paper No. 283*-0423.

Tables & Figures

Table 1: Descriptive characteristics of students and teachers

	Hand-selected comparison mean	School-grade comparison mean	Emergency licenses in district mean	Treatment mean
	(1)	(2)	(3)	(4)
Absence Sample, 2020-21				
Students				
Female	0.492	0.493	0.487	0.490
White	0.016	0.013	0.011	0.014
Black	0.963	0.954	0.954	0.965
Other Race/Ethnicity	0.022	0.033	0.035	0.021
Special Education Services	0.128	0.125	0.123	0.149
Title I	0.926	0.984	0.977	0.922
English Language Learner	0.012	0.019	0.022	0.010
2019-20 Absences	7.605	8.780	8.912	7.465
Missing 2019-20 Absences	0.111	0.093	0.092	0.130
N	*4243*	*183222*	*42419*	*4671*
Teachers				
Years Experience*	11.000	10.173	3.736	6.551
Years Teaching Experience*	10.288	10.052	3.673	6.042
Female	0.800	0.769	0.771	0.969
White	0.200	0.170	0.036	0.000
Black	0.800	0.814	0.959	1.000
Other Race/Ethnicity	0.000	0.016	0.005	0.000
N	*75*	*2671*	*617*	*98*
Test Score Sample, 2020-21				
Students				
Female	0.502	0.494	0.489	0.494
White	0.014	0.012	0.010	0.016
Black	0.972	0.955	0.955	0.967
Other Race/Ethnicity	0.015	0.033	0.035	0.017
Special Education Services	0.095	0.099	0.108	0.103
Title I	0.965	0.987	0.986	0.950
English Language Learner	0.008	0.020	0.023	0.005
2018-19 Test Score	-0.327	-0.323	-0.418	-0.347
Missing 2018-19 Test Score	0.231	0.299	0.285	0.244
N	*2182*	*47020*	*12227*	*2105*
Teachers				
Years Experience*	12.370	8.873	3.833	5.800
Years Teaching Experience*	11.407	8.779	3.800	4.967
Female	0.704	0.829	0.818	0.933
White	0.222	0.149	0.021	0.000
Black	0.778	0.843	0.979	1.000
Other Race/Ethnicity	0.000	0.008	0.000	0.000
N	*27*	*753*	*192*	*30*

Note: Columns present raw means. Test scores are standardized within grade, test, subject, and year using the statewide distribution such that the mean is 0 and the standard deviation is 1. Sample sizes are italicized.

* The years of experience variables are inconsistently documented and do not follow expected patterns, so I present them with caution. I do not include them in my main analyses.

Table 2: Student characteristics and pre-treatment balance

	Treatment vs. hand-selected	Treatment vs. school-grade	Treatment vs. emergency licenses
	(1)	(2)	(3)
Absence Sample, 2020-21			
Female	-0.002	-0.004	-0.001
White	-0.000	-0.003*	0.001
Black	0.001	0.006**	0.005
Other Race/Ethnicity	-0.000	-0.003*	-0.006
Special Education Services	0.017	0.006	0.015
Title I	-0.006	-0.001	0.001
English Language Learner	-0.001	-0.001	-0.005
2019-20 Absences	-0.131	0.098	-0.253
Missing 2019-20 Absences	0.029**	0.003	0.003
N	*8914*	*187893*	*47090*
Test Score Sample, 2020-21			
Female	0.012	-0.022	-0.005
White	0.007	-0.011*	0.001
Black	-0.010*	0.019**	0.004
Other Race/Ethnicity	0.004	-0.008**	-0.005
Special Education Services	0.011	-0.007	-0.015
Title I	0.000	-0.000	0.002
English Language Learner	-0.002	-0.001	-0.006*
2018-19 Test Score	-0.036	0.030	0.079
Missing 2018-19 Test Score	-0.095*	-0.003	0.031
N	*4287*	*49125*	*14332*

Note: Each cell in (1) through (3) presents a coefficient from a separate regression with the student characteristic as the outcome and a treatment dummy as the predictor of interest. Column 1 includes pair fixed effects, column 2 includes school-by-grade fixed effects (school-by-grade-by-subject for the test score analysis), and column 3 includes separate school and grade fixed effects. Standard errors are clustered at the teacher level. Sample sizes are italicized.

* p<0.05, ** p<0.01, *** p<0.001

Table 3: Treatment effects on absences and student achievement

	Hand-selected comparison	School-grade comparison	Emergency license comparison	Difference between treatment and emergency licenses in school-grade
	(1)	(2)	(3)	(4)
Number Absences, 2020-21	-0.720	0.029	0.243	-0.391
	(0.727)	(0.232)	(0.409)	[0.359]
	7784	*170578*	*41756*	*170578*
Number Absences, 2021-22	-0.692	0.234	0.315	-0.036
	(0.511)	(0.205)	(0.310)	[0.908]
	10731	*222405*	*41839*	*222405*
Number Absences, Pooled Years	-0.989**	0.136	0.931+	-0.107
	(0.371)	(0.152)	(0.504)	[0.829]
	18515	*392983*	*83595*	*392983*
Standardized Test Scores, 2020-21	0.001	0.007	-0.001	-0.018
	(0.052)	(0.051)	(0.044)	[0.631]
	4287	*49125*	*14332*	*49125*
Standardized Test Scores, 2021-22	-0.072	0.063	-0.046	-0.020
	(0.048)	(0.071)	(0.058)	[0.619]
	5958	*55329*	*12469*	*55329*
Standardized Test Scores, Pooled Years	-0.033	0.031	-0.045	-0.041
	(0.039)	(0.045)	(0.040)	[0.225]
	10245	*104454*	*26801*	*104454*

Note: Each cell in columns (1) through (3) reports a coefficient from a separate regression following equation (1). All test scores are standardized within grade, test, subject, and year using the statewide distribution such that the mean is 0 and the standard deviation is 1. All models control for student demographics, lagged absence rates, and lagged test scores. Column 1 includes pair fixed effects, column 2 includes school-by-grade fixed effects (school-by-grade-by-subject for the test score analysis), and column 3 includes separate school and grade fixed effects. Standard errors are clustered at the teacher level and presented in parentheses. Column (4) presents the difference in coefficients on a treatment dummy and an emergency license dummy in treated districts with school, grade, and

subject fixed effects, following equation (2). The associated p-value for the F-test of coefficient equivalency is in brackets. Sample sizes are italicized.
+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 4: Treatment effects on student achievement, separately by test subject

	Hand-selected comparison	School-grade comparison	Emergency license comparison	Difference between treatment and emergency licenses in school-grade
	(1)	(2)	(3)	(4)
ELA, 2020-21	0.125*	-0.113+	-0.034	-0.036
	(0.049)	(0.065)	(0.074)	[0.601]
	2384	*19212*	*5605*	*19212*
ELA, 2021-22	-0.101*	-0.016	-0.049	-0.043
	(0.041)	(0.085)	(0.039)	[0.408]
	3098	*22338*	*4837*	*22338*
ELA, Pooled Years	-0.004	-0.064	-0.025	-0.052
	(0.035)	(0.055)	(0.046)	[0.209]
	5482	*41550*	*10442*	*41550*
Math, 2020-21	0.097	0.208*	0.010	0.128
	(0.121)	(0.100)	(0.139)	[0.207]
	1003	*19175*	*5696*	*19175*
Math, 2021-22	-0.007	0.202*	0.178*	0.021
	(0.073)	(0.090)	(0.070)	[0.753]
	1844	*20491*	*5319*	*20491*
Math, Pooled Years	0.007	0.196**	-0.041	-0.033
	(0.055)	(0.068)	(0.099)	[0.657]
	2847	*39666*	*11015*	*39666*

Note: Each cell in columns (1) through (3) reports a coefficient from a separate regression following equation (1). All test scores are standardized within grade, test, subject, and year using the statewide distribution such that the mean is 0 and the standard deviation is 1. All models control for student demographics, lagged absence rates, and lagged test scores. Column 1 includes pair fixed effects, column 2 includes school-by-grade fixed effects, and column 3 includes separate school and grade fixed effects. Standard errors are clustered at the teacher level and presented in parentheses. Column (4) presents the difference in coefficients on a treatment dummy and an emergency license dummy in treated districts with school and grade fixed effects, following equation (2). The associated p-value for the F-test of coefficient equivalency is in brackets. Sample sizes are italicized.
+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 5: Differences in professional growth scores across teacher groups of interest

	Hand-selected comparison	School comparison	Emergency license comparison	Difference between treatment and emergency licenses in school-grade
	(1)	(2)	(3)	(4)
Summative Evaluation Score	-0.013	0.068+	0.119*	0.152***
	(0.065)	(0.036)	(0.049)	[0.000]
	180	*2521*	*521*	*2521*
Standard 1: Lesson aligned to standards	0.000	0.054	0.112+	0.153**
	(0.089)	(0.049)	(0.063)	[0.005]
	180	*2521*	*521*	*2521*
Standard 2: High levels of learning	-0.073	0.047	0.091	0.146*
	(0.108)	(0.055)	(0.076)	[0.017]
	180	*2521*	*521*	*2521*
Standard 3: Monitors student learning	0.061	0.074	0.169*	0.211***
	(0.108)	(0.056)	(0.076)	[0.001]
	180	*2521*	*521*	*2521*
Standard 4: Multiple ways for students to learn	-0.110	0.060	0.100	0.158**
	(0.108)	(0.054)	(0.069)	[0.008]
	180	*2520*	*521*	*2520*
Standard 5: Learning-focused classroom community	-0.012	0.143**	0.218**	0.243***
	(0.077)	(0.055)	(0.078)	[0.000]
	180	*2520*	*521*	*2520*
Standard 6: Classroom management	0.025	0.099	0.127+	0.181**
	(0.102)	(0.064)	(0.073)	[0.008]
	178	*2519*	*521*	*2519*
Standard 7: Classroom respect	-0.073	-0.032	0.004	-0.001
	(0.078)	(0.048)	(0.061)	[0.983]
	180	*2520*	*521*	*2520*
Standard 8: Professional learning	-0.012	0.106*	0.120+	0.158**
	(0.101)	(0.053)	(0.068)	[0.007]
	180	*2520*	*521*	*2520*
Standard 9: Effective communication with families	0.085	0.062	0.130*	0.116*
	(0.079)	(0.047)	(0.058)	[0.025]
	180	*2520*	*521*	*2520*

Note: Each cell in columns (1) through (3) reports a coefficient from a separate regression following equation (4). Column 1 includes pair fixed effects, columns 2 and 3 include school fixed effects. Standard errors are clustered at the teacher level and presented in parentheses. Column (4) presents the difference in coefficients on a treatment

dummy and an emergency license dummy in treated districts with school fixed effects, following equation (2). The associated p-value for the F-test of coefficient equivalency is in brackets. Sample sizes are italicized.
+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 6: Differences in retention rates across teacher groups of interest

	Hand-selected comparison	District comparison	Emergency license comparison	Difference between treatment and emergency licenses in school-grade
	(1)	(2)	(3)	(4)
Still teaching in a PBL district, 2021-22	-0.033	0.105*	0.081+	0.104*
	(0.051)	(0.044)	(0.048)	[0.024]
	190	*2792*	*718*	*2792*
Still teaching in a PBL district, 2022-23	0.020	0.102**	0.164***	0.133**
	(0.056)	(0.038)	(0.046)	[0.001]
	223	*3440*	*670*	*3440*

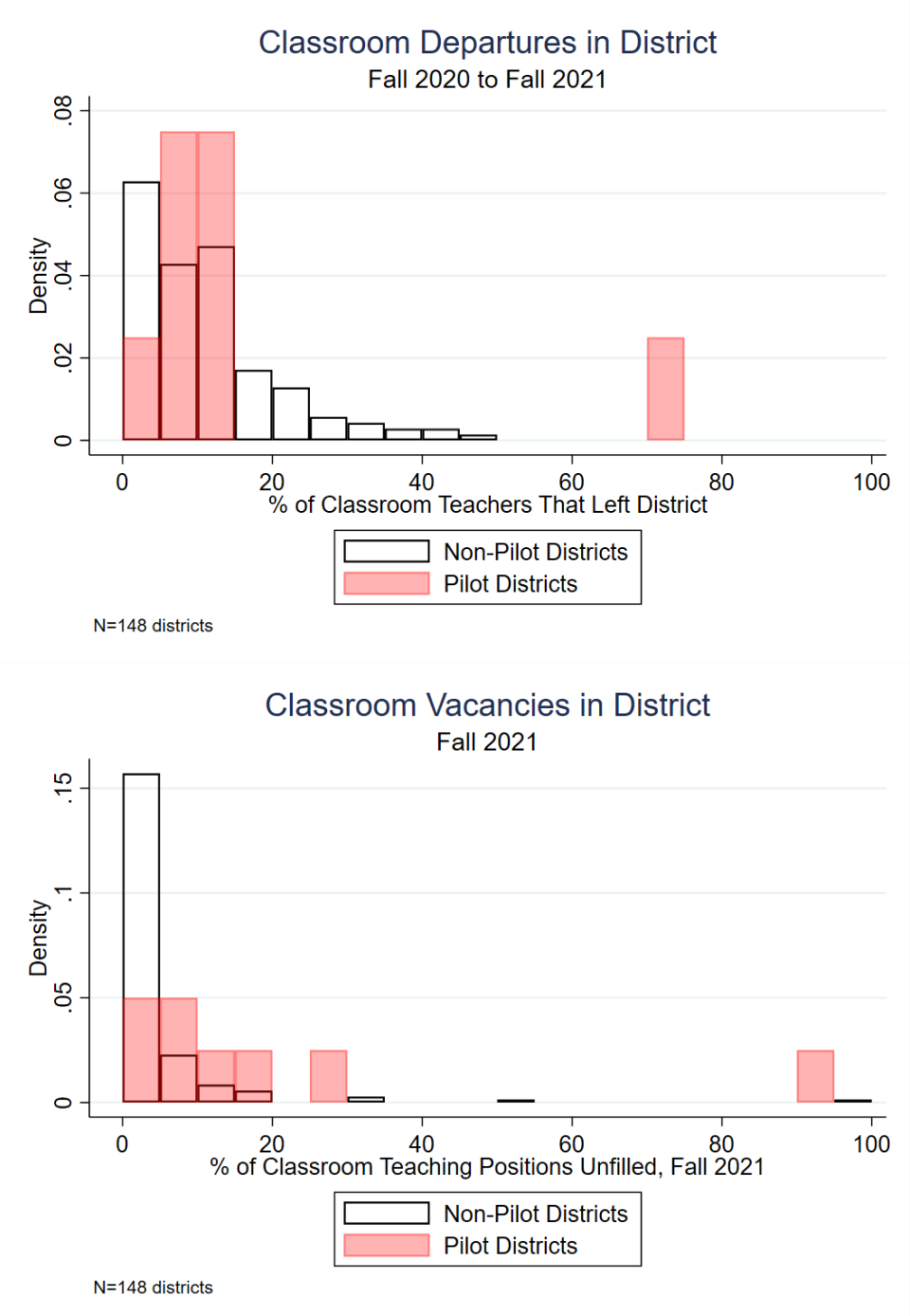
Note: Each cell in columns (1) through (3) reports a coefficient from a separate regression following equation (5). Column 1 includes pair fixed effects, and columns 2 and 3 include district fixed effects. Standard errors are presented in parentheses. Column (4) presents the difference in coefficients on a treatment dummy and an emergency license dummy in treated districts with district fixed effects. The associated p-value for the F-test of coefficient equivalency is in brackets. Sample sizes are italicized.
+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 7: Difference-in-difference estimates

	(1)	(2)	(3)
Pooled Subjects			
Treatment	0.013	0.013	-0.044
	(0.011)	(0.012)	(0.030)
	[-0.010, 0.035]	[-0.010, 0.035]	[-0.104, 0.016]
	212597	*212597*	*212597*
Math			
Treatment	0.025+	0.027+	-0.053
	(0.015)	(0.015)	(0.039)
	[-0.004, 0.054]	[-0.002, 0.056]	[-0.130, 0.023]
	106235	*106235*	*106235*
ELA			
Treatment	0.001	-0.001	-0.034
	(0.017)	(0.018)	(0.046)
	[-0.033, 0.035]	[-0.036, 0.033]	[-0.125, 0.057]
	106362	*106362*	*106362*
Group Fixed Effects	School-Grade	School-Grade	School-Grade
Time Fixed Effects	Year	Grade-Year	Grade-Year
Treatment	Binary	Binary	Proportion

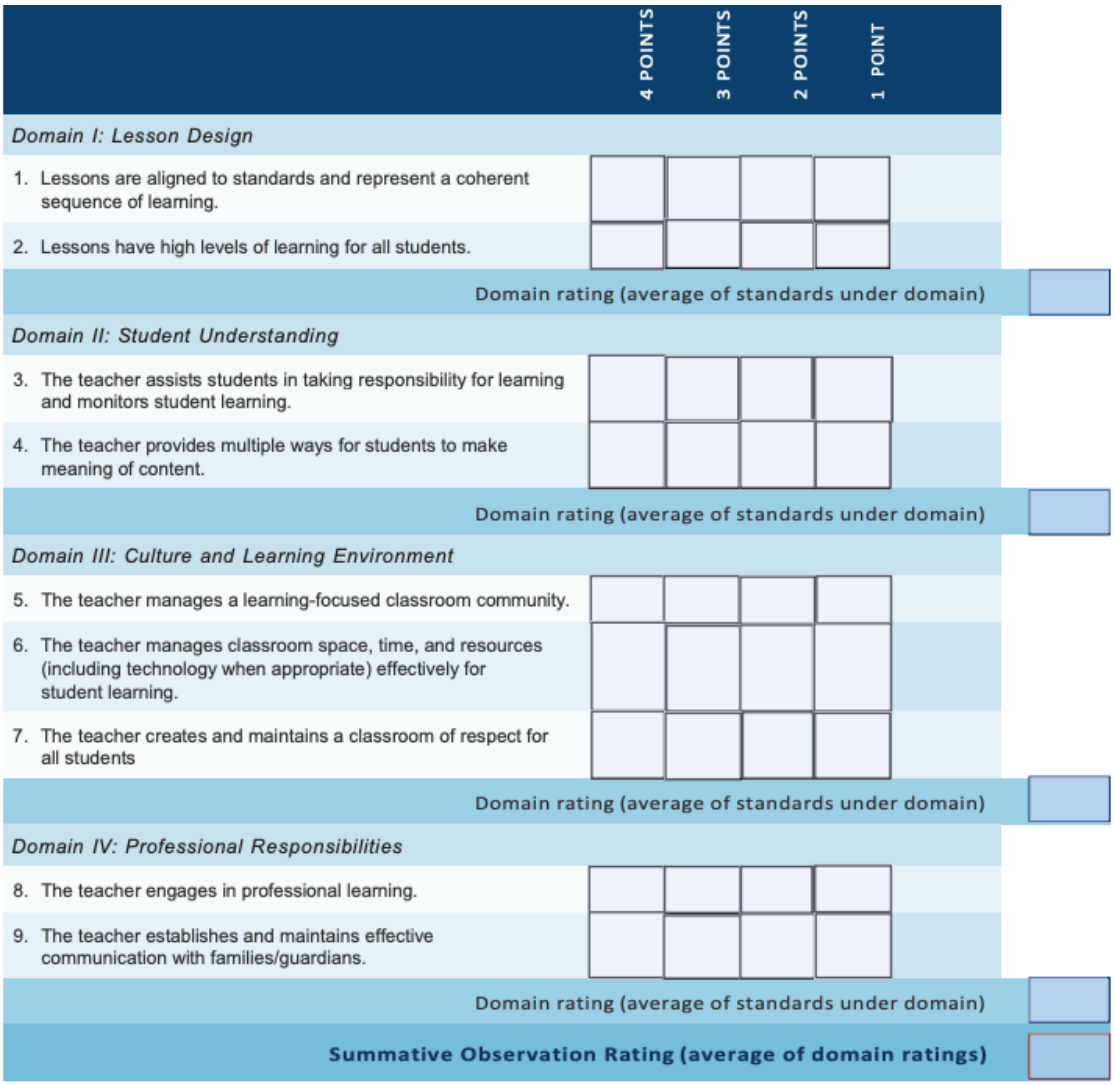
Note: Each cell in columns (1) through (3) reports a coefficient from a separate regression following equation (6). All models include fixed effects for test subject and school-grade cells and are limited to treated districts. Column (1) includes year fixed effects and columns (2) and (3) include grade-by-year fixed effects. The treatment variable is binary in columns (1) and (2) and a proportion in column (3). Robust standard errors are presented in parentheses. Confidence intervals are in brackets. Sample sizes are italicized.
+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Figure 1: Reported teaching vacancies in district surveys



Note: Histograms of responses on state survey to all districts on teacher shortages. Bins are 5 percentage points wide.

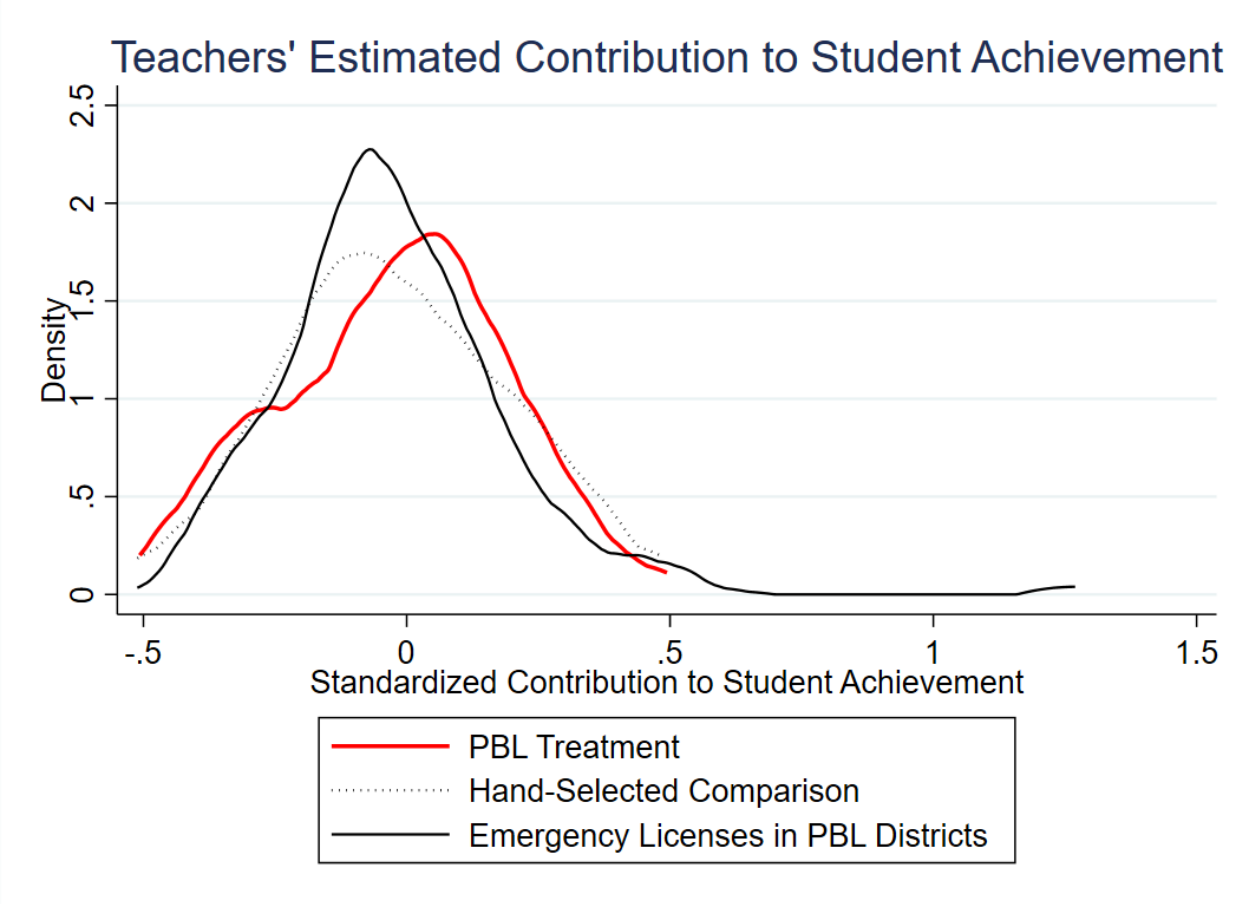
Figure 2: Teacher observation rubric



	4 POINTS	3 POINTS	2 POINTS	1 POINT	
Domain I: Lesson Design					
1. Lessons are aligned to standards and represent a coherent sequence of learning.					
2. Lessons have high levels of learning for all students.					
Domain rating (average of standards under domain)					
Domain II: Student Understanding					
3. The teacher assists students in taking responsibility for learning and monitors student learning.					
4. The teacher provides multiple ways for students to make meaning of content.					
Domain rating (average of standards under domain)					
Domain III: Culture and Learning Environment					
5. The teacher manages a learning-focused classroom community.					
6. The teacher manages classroom space, time, and resources (including technology when appropriate) effectively for student learning.					
7. The teacher creates and maintains a classroom of respect for all students					
Domain rating (average of standards under domain)					
Domain IV: Professional Responsibilities					
8. The teacher engages in professional learning.					
9. The teacher establishes and maintains effective communication with families/guardians.					
Domain rating (average of standards under domain)					
Summative Observation Rating (average of domain ratings)					

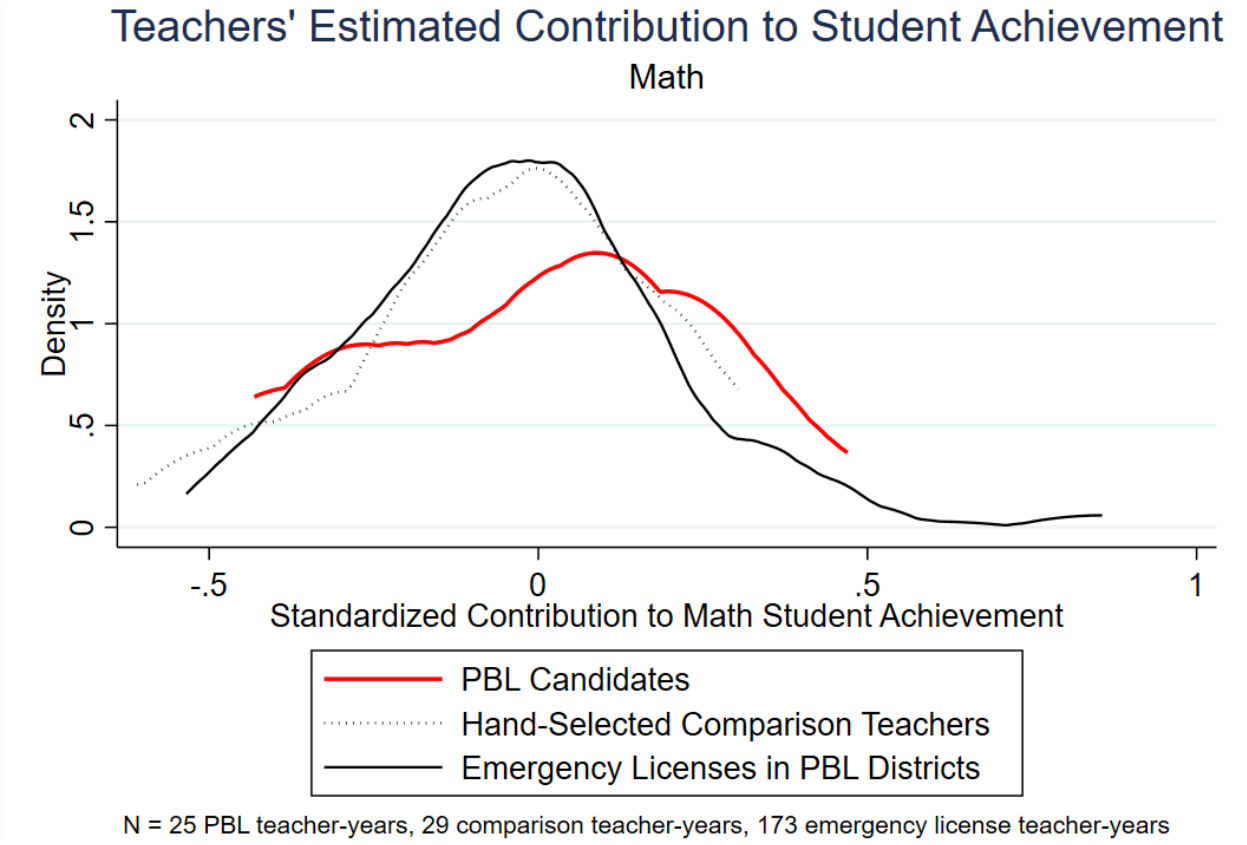
Note: Retrieved from https://www.mdek12.org/OEE/Teacher

Figure 3: Estimated value-added across teacher groups of interest: pooled subjects



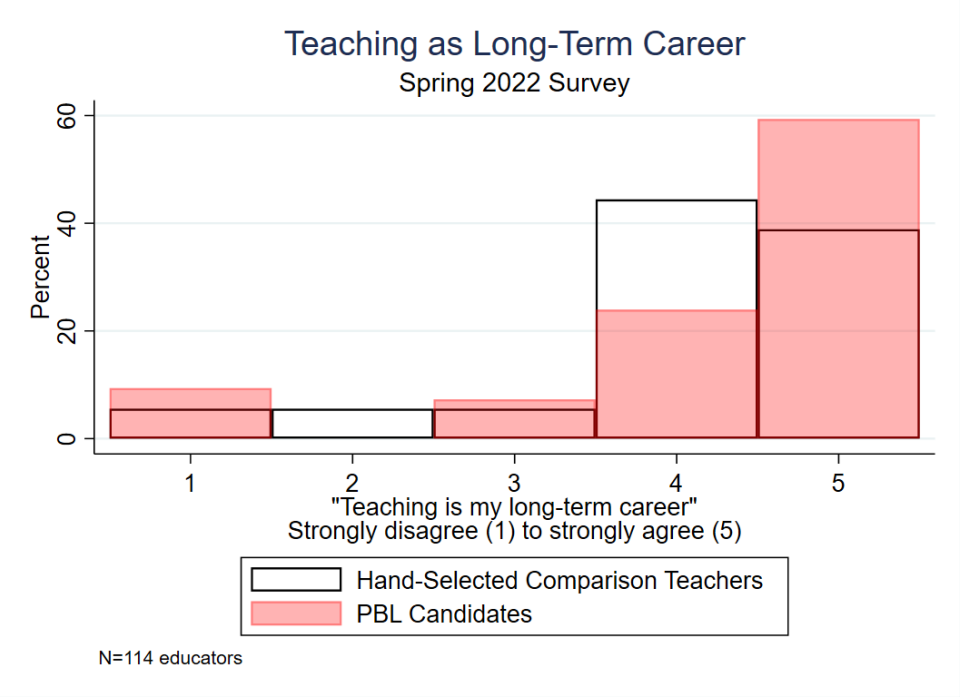
Note: Distribution of $\widehat{\alpha_{jt}}$ estimated following equation (3). Estimation is limited to pilot districts and includes grade, subject, and district fixed effects as well as controls for student demographics and lagged test scores.

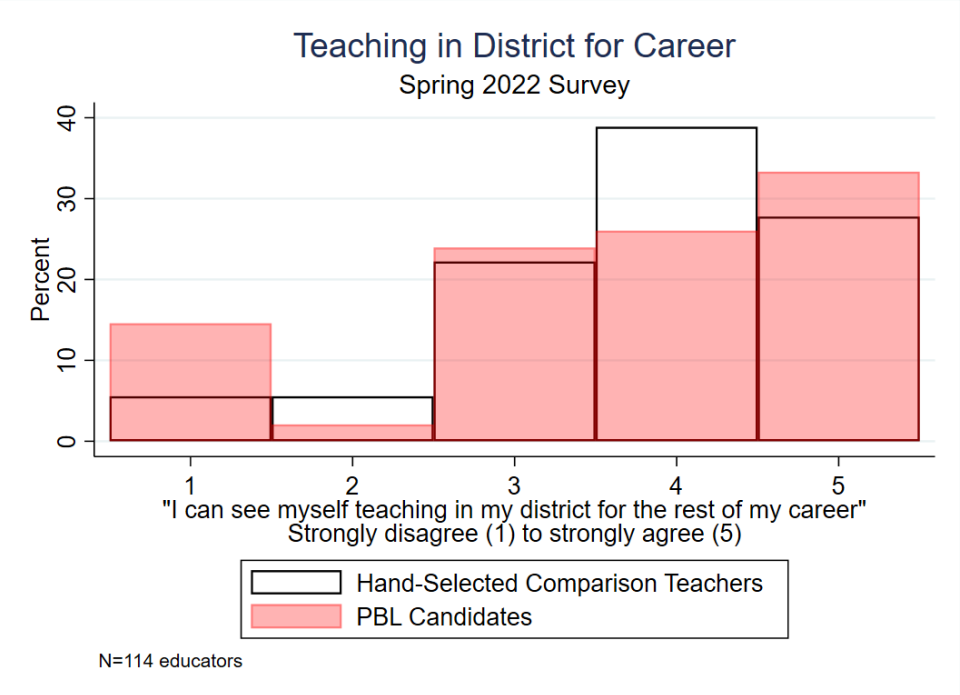
Figure 4: Estimated value-added across teacher groups of interest: math



Note: Distribution of $\widehat{\alpha_{jt}}$ estimated following equation (3). Estimation is limited to pilot districts and includes grade and district fixed effects as well as controls for student demographics and lagged test scores.

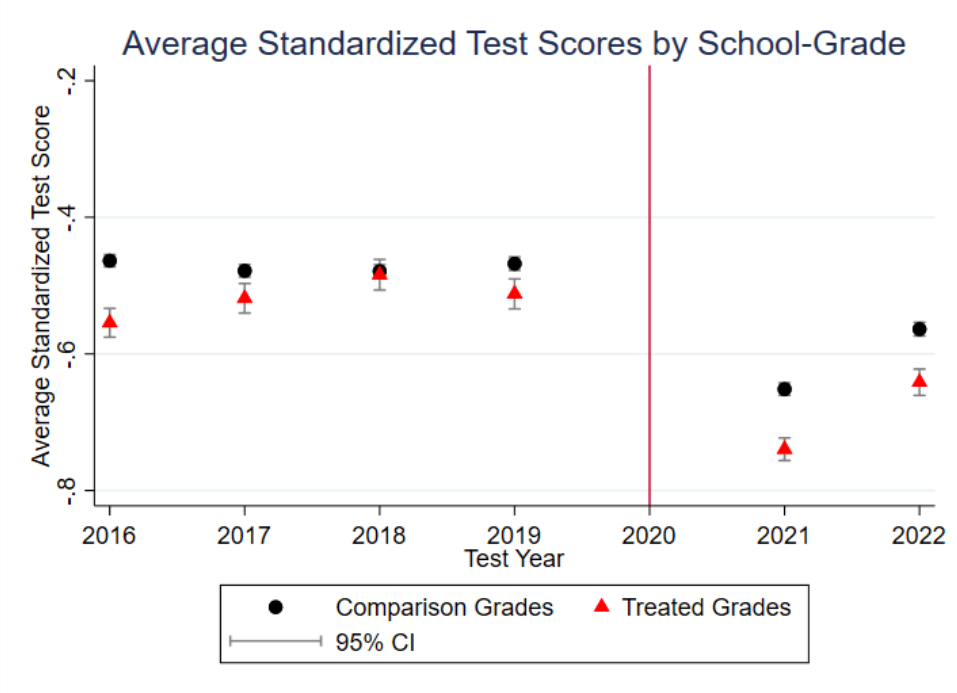
Figure 5: Teacher career interests, from program survey

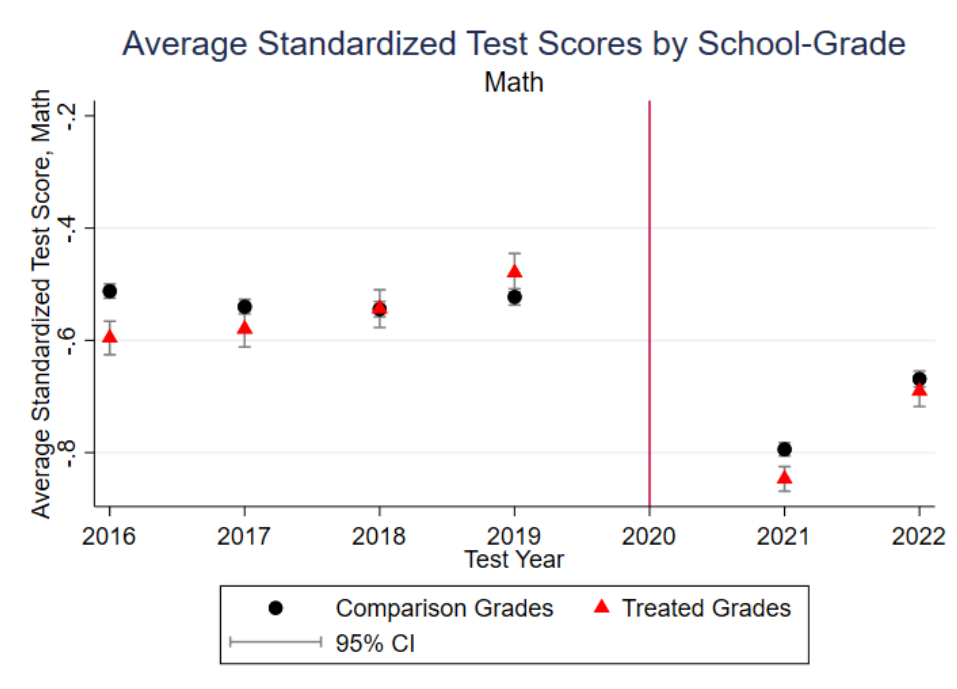




Note: Histograms of responses from state survey to all PBL candidates and hand-selected comparison teachers. Each question has a Likert scale response range: strongly disagree (1), somewhat disagree (2), neutral (3), somewhat agree (4), strongly agree (5).

Figure 6: School-grade-year-level mean test scores, pooled (panel A) and math (panel B)





Note: Means of standardized test scores of all students in treated and comparison school-grades in each year and the associated 95% confidence interval for each mean. Test scores were standardized using the statewide distribution. Estimates limited to treated districts. No tests were administered in 2020 due to the COVID-19 pandemic.