Performance Pay and Educational Inequality: Evidence from Texas

Sebastián Arechaga*
University of California, Berkeley

Job Market Paper

This version: November 27, 2025

Download Latest Version

Abstract

Performance-based incentives in the public sector often raise concerns that they exacerbate inequality by disproportionately benefiting already advantaged localities. This tension is especially salient in education, where teachers shape student achievement, yet high-need schools struggle to retain and motivate talent. I study the *Teacher Incentive Allotment* (TIA), a statewide performance-pay reform launched in Texas in 2019. The program provides substantial bonuses to highly effective teachers, with larger amounts for schools serving more vulnerable students. Exploiting the staggered rollout of TIA across nearly 400 districts and 3,000 schools, I estimate its causal effects using an event-study design. TIA raised standardized test scores by 0.04 standard deviations on average, with the largest gains in vulnerable schools, reducing within-district inequality by 2.3–8.7%. Teacher turnover declined temporarily, creating a more experienced and stable workforce. Among movers, designated teachers were more likely to transfer to high-vulnerability schools, where bonuses were higher. Observable changes in teacher composition explain at most 14% of gains, implying that most improvements stem from higher instructional effectiveness. By disproportionately boosting achievement in disadvantaged schools, TIA narrowed educational gaps, showing that well-targeted incentives can enhance both efficiency and equality in education.

*sebastian_arechaga@berkeley.edu

I thank Guo Xu, Ernesto Dal Bó, and Reed Walker for their extensive advice and support. For helpful comments and suggestions, I thank Pía Arce, Morgan Foy, Fred Finan, Jesse Rothstein, Cailin Slattery, Kelsey Jack, Hilary Hoynes, Ricardo Perez-Truglia, Steven Raphael, and many seminar participants. I also thank staff at the Texas Education Agency for their diligent help in accessing the data.

1 Introduction

The delivery of high-quality public services remains a central challenge in public administration. Outputs are often difficult to measure, oversight mechanisms are weak, and rigid pay structures limit incentives for effort (Bertrand et al., 2020). In addition, attracting and retaining talented workers in such contexts is difficult, as small financial incentives rarely offset poor infrastructure and amenities (Dal Bó et al., 2013). Education exemplifies these broader challenges. Teachers play a central role in student learning (Chetty et al., 2014a,b), yet many U.S. schools, especially those serving disadvantaged populations, struggle to attract and retain effective educators (Nguyen et al., 2020; García and Weiss, 2019).

Despite the importance of these challenges, little is known about how to attract and motivate talented workers in vulnerable or remote places. Texas's recent policy reforms address this issue directly. In 2019, the state launched the *Teacher Incentive Allotment* (TIA), a statewide initiative that provides substantial performance-based bonuses to highly effective teachers. The program seeks to improve teacher performance and retention, particularly in high-need and rural schools, by linking bonus amounts to school vulnerability levels: teachers in more disadvantaged settings receive higher awards, aligning financial incentives with the state's equity and performance goals. The Teacher Incentive Allotment Program thus offers a valuable setting to study the economics of public service delivery, specifically, how financial incentives shape student achievement, teacher labor decisions, and the distribution of educational opportunity across schools.

The rollout of TIA offers a unique empirical opportunity: the program was implemented gradually across 376 independent school districts (out of 1,012 school districts in Texas) and over 3,000 schools. I exploit this staggered adoption in a quasi-experimental, event-study design. Rich administrative data, including school-by-grade-by-year test scores and teacher-level records on employment, enable me to track the average effect of implementing the program and the impact across schools with different levels of vulnerability indicators. In addition, information about teachers' rewards allows me to study how designated teachers respond to incentives. This event study framework allows for credible causal inference on TIA's effects by using later adopters and non-adopters as counterfactuals.

I find that performance incentives improved student achievement. On average, TIA increased standardized test scores by about 0.04 standard deviations in its first few years. These gains built up over one to two years of exposure, suggesting that the impact grows as the program matures. In addition, I observe positive changes in teacher retention and the composition of schools' workforce. Teacher turnover declined in TIA districts, leading to a more experienced and stable teacher workforce on average. Beyond the average effects, I explore the mechanisms driving the improvement in student achievement. The evidence suggests that a portion of the gains arises from a selection effect -schools retaining and attracting more experienced or effective teachers. However, this channel explains at most 14 % of the overall impact, indicating that the majority of the effect stems from productivity enhancements, such as greater teacher effort and improvements in instructional or managerial practices.

High-powered incentives often risk widening inequalities: when rewards depend on performance, better-resourced schools and more advantaged students are typically positioned to gain the most. The

Teacher Incentive Allotment (TIA) was designed with this challenge in mind, by channeling additional resources to high-vulnerability schools, the program was explicitly designed to reduce inequality in student achievement. I provide new evidence on this dimension by comparing achievement gains between the most disadvantaged and most advantaged schools under TIA. The results show that the gains were concentrated in schools with the highest levels of vulnerability. In addition, if districts differ in unobserved ways—such as changes in administrative quality, leadership turnover, or other changes at the local governance level—that could jointly affect both the decision to adopt the program and its effectiveness. Controlling for district-by-year fixed effects absorbs all time-varying shocks common to schools within the same school district. The specification rules out the possibility that these district-level factors drive the results; when controlling for district-by-year fixed effects, the contrast becomes even more pronounced, suggesting that the gains in equality are driven by within-district differences in the vulnerability level of schools.

The reduction in the achievement gap is substantial. The largest gains were observed in schools serving more vulnerable populations. These improvements translated into a decrease in within-district inequality —across conventional measures— of approximately 2.3 to 8.7 %. The reduction in educational disparities is even more pronounced across schools with varying vulnerability levels. In adopting districts, the slope of the relationship between average test scores and the vulnerability index is -0.378, implying that more vulnerable schools perform markedly worse on standardized tests. This means that two schools differing by one point on the *vulnerability score*—roughly equivalent to a 25-percentile difference in vulnerability—were expected to differ by 0.378 standard deviations in average test scores, with the more vulnerable school performing worse. After accounting for potential confounders, my estimates indicate that TIA attenuated this negative relationship by 0.036 to 0.081 points, depending on the specification. These results correspond to a 9% to 21% reduction in the achievement gap as measured by the correlation between test scores and school vulnerability.

Finally, despite an overall increase in teacher retention, designated teachers tended to gravitate toward high-need campuses. After being recognized as top performers and receiving bonuses, they became more likely to move to schools with higher vulnerability ratings, where the financial rewards are greater. This selective migration of talent indicates that the policy effectively incentivized teachers to reallocate toward students who stand to benefit most. Moreover, this dynamic helps explain why the average effects estimated in this paper capture short-run impacts. In the long run, the remaining effects of the policy would operate primarily through changes in teacher effort, instructional practices, new teacher entry, and a reduction in teachers leaving the workforce, rather than through further reallocation across schools.

These findings are consistent with the view that financial incentives can motivate teachers to remain in challenging environments and to focus more intensively on raising student achievement. In addition, these findings are important for policy: they suggest that well-designed teacher incentives can enhance student achievement and reduce educational differences by motivating and retaining talented teachers, especially in the schools that have struggled the most.

1.1 Related Work

These findings contribute to several areas of research. First, they add to the literature on performance-based incentives for teachers. Most evidence on teacher incentives comes from contexts where the main constraints on learning stem from limited resources and low baseline achievement—for instance, Glewwe et al. (2010) in Kenya, Muralidharan and Sundararaman (2011) in India, Mbiti et al. (2019) in Tanzania, Barrera-Osorio and Raju (2017) in Pakistan, and Leaver et al. (2021) in Rwanda. In contrast, evidence from systems where the central challenges involve sustaining effort, accountability, and motivation within mature institutions is more limited and shows mixed results. Fryer (2013) studies an experiment in New York City and finds no average effect on achievement. Goodman and Turner (2013) show that the program led to small increases in achievement when teachers have incentives to free ride —a result often cited to underscore the importance of incentive design. Yet this program involved relatively modest bonuses (up to $3,000), and few studies have evaluated large-scale performance-pay reforms offering substantial awards ranging from $3,000 to $32,000. Other U.S. studies, such as Springer et al. (2011)'s evaluation of Nashville's program, which rewarded between $5,000 and $15,000, also found null effects. Biasi (2021) shows results of a reform occurring at the state level in Wisconsin where school districts gained discretion over their compensation structures. Biasi (2021) show that the policy lead to improvements on student achievement. In Texas, districts already had the ability to define their pay structures prior to the introduction of TIA. Under TIA, districts can tailor elements of their teacher evaluation systems, but all modifications operate within a performance-based pay framework established by the state.

This study differs in both scope and design. It examines the implementation of a statewide program spanning roughly 400 school districts in Texas—making it one of the largest performance-pay reforms evaluated to date. While an early evaluation of the Teacher Incentive Allotment (TIA) using synthetic control methods finds large achievement gains among the first 19 adopters (Kirksey et al., 2024), this paper differs from that analysis in several ways: (i) using more granular data allows me to control for differences in school-level characteristics; (ii) exploiting staggered adoption across nearly 400 districts instead of focusing solely on early adopters, allowing for a more representative assessment of the program's long-run effects; (iii) finding smaller but sustained achievement gains of about 0.04 standard deviations that build over one to two years, compared to 0.15 s.d. in the initial study; (iv) incorporating teacher-level data to analyze within-district heterogeneity in teacher behavior and mobility; and (v) provide relevant findings on the effect of the policy on educational inequality, with an emphasis on how this result is achieved through the design of the policy.

The TIA's distributional structure motivates the paper's second contribution, which relates to the retention of workers in vulnerable or hard-to-staff areas. Dal Bó et al. (2013) study how financial incentives shape recruitment and find that, while higher wages attract more able applicants, distance and poor local amenities substantially reduce acceptance rates. Similarly, Falch (2011) finds that a 10% wage increase lowers the probability of voluntary quits in hard-to-staff schools in Norway. Burgess et al. (2022) find that schools increase teacher retention after a reform that rereplacesentralized pay scales with a performance pay regime. Biasi (2021) shows that, in Wisconsin, higher-quality teachers moved to districts that adopted flexible pay structures linked to merit or performance. I contribute to

this literature by showing that allocating larger performance-based payments to disadvantaged schools significantly reduces teacher turnover. Moreover, because bonus amounts are recalculated when teachers move between schools with different levels of vulnerability, I find that high-performing teachers are more likely to relocate to schools with higher levels of vulnerability.

Third, I contribute to the literature on reducing inequality in the provision of public services. For instance, it remains an open question whether performance-pay schemes affects educational equality (Biasi et al., 2021). On one hand, teachers in more vulnerable communities may generate higher returns to effort because they start from a lower baseline and have stronger incentives to qualify for the bonus (Muralidharan and Sundararaman, 2011). On the other hand, they often face binding constraints—such as limited resources, weaker management, or community instability—that may limit their ability to respond to incentives (Rasul and Rogger, 2018). In this paper, I find that performance-based bonuses have significantly larger effects on student achievement in vulnerable schools. These schools receive larger financial incentives, and the bonuses are financed with state funds, suggesting that the design of the program plays a crucial role for equality.

Finally, my work relates to the literature examining the mechanisms through which performance-pay programs operate. A particularly relevant study is the two-tier experiment in Rwanda's primary schools by Leaver et al. (2021), which separately identified the effects of incentives on selection - who enters and stays in the schools - and effort. By the second year, students taught under performance contracts scored about 0.2 standard deviations higher than those in the control group, with roughly one-fifth of the gain driven by improved teacher selection and the remainder by increased effort. In my analysis, I find a comparable share of the overall effect attributable to selection, suggesting that similar mechanisms may be at play even in a high-income setting.

The rest of the paper proceeds as follows. Section 2 describes the institutional background of the TIA, outlining its implementation timeline, design, and the variation in incentive amounts across schools. Section 3 introduces the data and provides descriptive evidence comparing adopting and non-adopting districts. Section 4 details the identification strategy and presents the main results on student achievement, teacher composition, and retention. Here, I also examine the mechanisms behind the observed achievement gains, focusing on whether they arise from changes in the composition of the teacher workforce—by attracting and retaining better teachers—or from improvements in effort and instructional practices. Section 5 analyzes the implications of the design of the policy on educational inequality, emphasizing heterogeneity by vulnerability level, which is closely linked to bonus size, and within school districts measures of inequality. In addition, here I explore whether financial incentives encourage high-performing teachers to move to more disadvantaged schools. Section 7 concludes.

2 The Teacher Incentive Allotment program

The Texas Legislature established the Teacher Incentive Allotment (TIA) through House Bill 3 (HB 3) during the 86th Texas Legislative Session in 2019. This bill was passed and signed into law in June 2019. The bill was championed by Republican leadership but received broad support across party lines due

to its emphasis on increased funding for public education, teacher compensation, and student outcomes with the new Teacher Incentive Allotment program (TIA).

TIA intended to attract and retain highly effective teachers, especially in rural and high-need schools. In 2019, the first cohort of school districts implemented the program, and the first *"designated teachers"* received their allotment in 2020. *"Designated teachers"* is the name received by teachers who, according to the metrics established by the districts, are high-performing teachers who deserve a bonus on top of the salary schedule defined by the school district.

High-performing teachers can fall into three designation categories based on their evaluation results: Recognized (the lowest designation level), Exemplary, or Master (the highest designation level). Teachers retain their designation for five school years. Designations may be revoked if the teacher is sanctioned in a way that suspends the teaching certificate or if the Texas Education Agency determines that the designation was improperly issued.[1] During the five-year period, the designation is portable across campuses and districts. A teacher may be re-submitted for a higher designation before the original five-year term expires; if the teacher is elevated (for example, from Recognized to Exemplary), the new designation replaces the previous one and restarts the five-year clock. After five years, the designation expires unless the district submits updated performance evidence that qualifies the teacher for renewal. Importantly, the designation does not automatically adjust downward if subsequent performance falls. This structure could, in principle, reduce effort after a designation is obtained. At the same time, teachers holding Recognized or Exemplary status face incentives to increase effort in order to move up to a higher-paying designation. The net effect on sustained effort is therefore ambiguous ex ante.

Adoption

Participation in the Teacher Incentive Allotment program is voluntary for Texas school districts, and a process of several steps is required to join the program to ensure that each district can rigorously identify high-performing teachers. First, districts must design their own *"local designation system"* adjusted to their needs. The system must combine teacher observation scores and measurements of student academic growth. Some districts also choose to incorporate other indicators, such as attendance. The local system can be tailored to the district's context, but it must meet standards pre-established by the state. Because adoption is voluntary, adopting districts may differ from non-adopters in ways that could bias the estimates. I address this concern using an event-study design, for which identification requires that treated and untreated districts have followed parallel trends in the absence of the program. I assess the plausibility of this assumption by examining pre-treatment dynamics and find no evidence of differential trends before implementation, supporting the validity of this assumption.

The entire approval process can take one to two years. Each year, applying districts prepare their systems and submit them for approval. After approval, the district can submit data on its teachers for consideration for TIA designation.

[1] For example, if the teacher is listed in the Do Not Hire registry, the issuing district's designation system is revoked, or an irregularity in the designation process is identified.

Teachers' eligibility and designation

Only teachers who are assigned to campuses that have been submitted and approved as part of the district's designation plan are eligible for consideration, and the Texas Educational Agency must approve that list. Administrative data from the TIA website indicates that all districts participating in the program had included all schools. Eligible teachers must hold a valid Texas teaching certification, be employed full-time, and have an assignment that contributes to student credit under the state's PEIMS reporting system, which usually means teachers of core subjects. Teachers do not need to submit an individual application to be considered for a designation. Their inclusion is automatic if they work in an eligible school and meet the basic criteria.

Teachers' allotment

The state finances the cost of designated teachers, but the funds generated by each designation are not automatically transferred to that teacher. Instead, for each designated teacher the TIA allocates funds to the teacher's district. Districts are required to spend at least 90% of this allotment on teacher compensation. However, they have discretion over how to distribute the allotment, for example, whether to direct the entire amount to the designated teacher or to share a portion with other staff. In practice, most districts allocate roughly 90% of the funds associated with a given designation to the designated teacher.

The amount generated by each designated teacher depends on two key factors: a *"school-level score"* and the designation level of the teacher. The *"school-level score"* is a score that combines the socioeconomic tier of the school— based on the share of economically disadvantaged students —with the rural status of the school. I refer to this school-level combination of socioeconomic tier and rural status as a *"vulnerability score"*, since it reflects the degree of contextual disadvantage due to economic disadvantage and geographic isolation. The second component is the designation level of the teacher, which depends on the performance of the teacher. A teacher awarded with a "recognized" level may generate between \$3,000 and \$9,000 USD in additional allotment depending on the school's "vulnerability" level, "exemplary" teachers between \$6,000 and \$18,000 USD, and "master" teachers between \$12,000 and \$32,000 USD.[2] These amounts are sizable relative to the average base pay of teachers in the sample, which was slightly above \$55,000 in 2018.

When a designated teacher moves to a different district, the designation remains valid and the teacher continues to generate TIA funding. However, the amount associated with that designation may change, because TIA allotments are a function of the *"vulnerability score"* of the school. As a result, the same teacher can generate different funding amounts in different placements. Importantly, the designation continues to generate funding even if the teacher moves to a district that has not adopted TIA, so teachers do not forfeit the financial benefit of their designation by transferring to a non-adopting district.

The allotment formula for each designation level is straightforward: it increases linearly with the school's *vulnerability score*. This score is a continuous index ranging from 0 (least vulnerable) to 4 (most vulnerable), constructed by the Texas Education Agency using information on students' socioeconomic

[2]Appendix A contains an example with a detailed explanation of the designation process

composition and the school's rurality. The distribution of vulnerability scores, shown in Figure 1, indicates that the distribution is skewed toward lower scores (less vulnerability). The allotment formulas for each designation level are presented in Table 1. Both the base payment and the multiplier grow with the designation rank—being highest for *Master* teachers, followed by *Exemplary*, and then *Recognized* teachers.

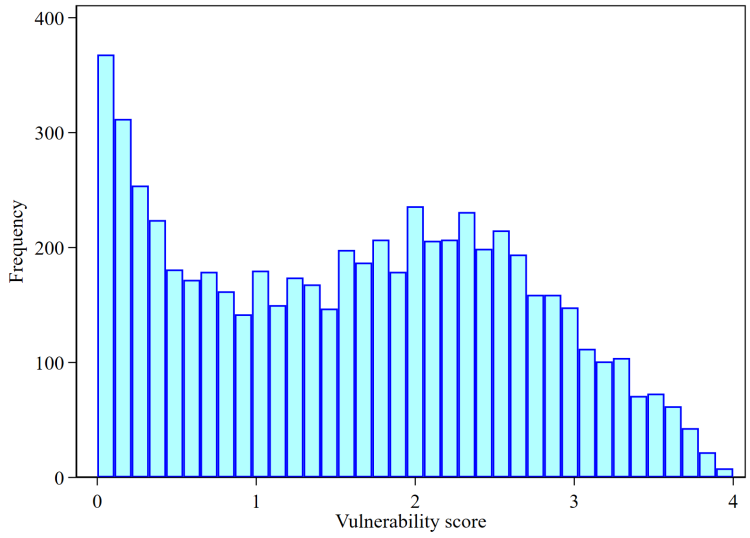
The magnitude of these difference on financial incentives is substantial. For instance, a teacher with a *Master* designation in the least vulnerable school receives an allotment of $12,000, while the same designation in the most vulnerable school yields $32,000. The steep gradient in allotment by school vulnerability implies that designated teachers face substantial incentives to transfer to higher-need schools, where their designation yields greater financial rewards. Across designation levels, the difference between a *Recognized* and an *Exemplary* teacher ranges from roughly $3,000 in low-vulnerability schools to $9,000 in high-vulnerability ones, and the gap between an *Exemplary* and a *Master* teacher increases from about $6,000 to $14,000. These represent sizable supplements to the average teacher base salary of around $55,000 in the sample, highlighting the strong financial incentives embedded in the program's design.

Table 1: Allotment Formula for Designated Teacher by Designation Level

Recognized	\$3,000+\$1,500 $\times$ *school's vulnerability score*
Exemplary	\$6,000+\$3,000 $\times$ *school's vulnerability score*
Master	\$12,000+\$5,000 $\times$ *school's vulnerability score*

Source: https://tiatexas.org/

Figure 1: Number of Teachers by Adoption and Designation Status.



The graph shows the number of teachers in schools by the level of vulnerability of the school using teacher-level observations for 2018 linked to schools' level of vulnerability. The list of teachers and schools' level of vulnerability were requested to Texas Education Agency and provided under Freedom of Information Act.

3 Data and Descriptive Statistics

3.1 Data

This study leverages administrative data from the Texas Education Agency (TEA). The datasets cover the universe of Texas public schools and school districts from the 2010–2011 through 2023–2024 school years.

I use student performance data from the State of Texas Assessments of Academic Readiness (STAAR) test at the school-by-grade-by-subject level. To construct school-level data on teacher retention and inter-district mobility, I use teacher-by-year personnel records from Texas's PEIMS database, obtained through a Freedom of Information Act request. This data includes teachers' identifiers, teachers' schools and school district, their demographics, qualifications, experience, and yearly school assignments. These records are linked to data indicating which districts adopted the program, which teachers received TIA designations each year, their designation level, and a school-level vulnerability index used to define the bonus schedule for designated teachers.

Finally, I incorporate contextual variables. The primary contextual variables relate to measures of the impact of COVID-19 over the years 2020 and 2021. To control for differential trends caused by COVID, I need to control for variables that are closely associated with the implications of COVID for education. This is the reason why I use the number of weeks spent in hybrid and remote teaching modes during 2020 at the district level. Some big districts had performance pay systems in years before TIA; those school districts were removed from this analysis.[3]

3.2 Adoption

3.2.1 Rollout

The adoption of TIA by school districts is voluntary. Therefore, cross-district variation in TIA implementation is not exogenous. Understanding the evolution of TIA adoption and designation across time is essential for interpreting subsequent analysis results under the required assumptions for identification. In addition, it helps us to understand which districts are more likely to implement such policies.

To begin, I first examine descriptive trends in the program rollout. Figure 2a visualizes the number of teachers in TIA-participating versus non-participating districts over time. The number of teachers in school districts with TIA has increased substantially over time, with about 120,000 teachers in TIA districts. This represents only about 30% of the total number of teachers in the state. I do not have teachers' data available for 2024, but using the number of teachers in each school district in 2023 and the number of districts adopting the program in 2024, the number of teachers under a school district with TIA should have increased by over 50%.

Figure 2b plots the number of designated teachers by designation level from 2020—when teachers first began receiving designations—through 2023. In 2020, approximately 18% of teachers in TIA districts

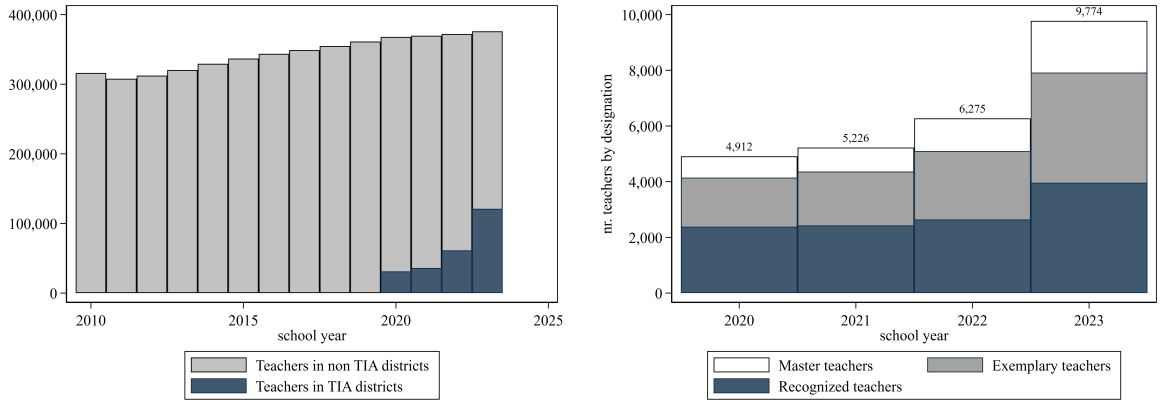
[3]The school districts with prior performance pay programs are: Spring ISD, San Antonio ISD, Dallas ISD, Houston ISD, and Austin ISD. All these districts had their own performance pay program for at least one year between 2005 and 2019. Some of them included a few schools from the district.

held a designation, but by 2023 this share had declined to about 8% (9,774 out of roughly 120,000 teachers). This pattern suggests that while the TIA's reach expanded across districts over time, the designations—and corresponding monetary allotments—remained concentrated among a relatively small group of high-performing teachers.

The share of teachers directly benefiting from the program varies considerably across adopting districts and over time, meaning that most teachers did not immediately experience direct financial gains. This has several implications. First, if the program motivates teachers to increase effort, many may improve performance without necessarily reaching the threshold for designation, leading to achievement gains without proportional increases in program spending. Second, teachers' behavioral responses depend not only on the size of the potential bonus but also on their perceived probability of earning it. When relatively few teachers receive designations, the expected reward from additional effort is lower, potentially dampening the incentive effect. In sum, teachers' responses to the policy depend on the magnitude of the monetary rewards and their expected likelihood of receiving them.

Figure 2: Number of Teachers by Adoption and Designation Status.

(a) Nr. of Teachers in TIA and non TIA Districts. (b) Nr. of Teachers in TIA Districts by Designation.



The first graph shows, in gray, the total number of teachers in public schools in Texas each year. The blue bar shows the number of teachers under a TIA regime. The second graph shows the number of teachers who receive an allotment because they obtained a designation, the blue bars show the number of teachers in the "*recognized*" category, the gray bars show the number of teachers in the "*exemplary*" category, and the white bars with black contour show the number of teachers in the "*master*" category (the highest category of designation). The list of teachers and designated teachers were requested to Texas Education Agency and provided under Freedom of Information Act.

3.2.2 Adopters and Non-Adopters

The effects estimated in further sections will take advantage of the fact that there are district adopting the policy on different years. To understand what types of districts adopt the reform, I answer the following question: *How do adopters and non-adopters compare with each other?* The adoption of TIA occurred in distinct waves between 2019 and 2024, corresponding to cohorts of districts approved by the state each year. Table 2 presents descriptive statistics for non-adopters ("Control"), and the difference between non-adopters and adopters by adoption wave. I divide adopters in two groups early-adopters –school districts adopting the program between 2019 and 2021– and late adopters –school districts adopting the program between 2022 and 2024. The table shows systematic differences between early adopters, late adopters, and non-adopters.

Academically, early-adopting districts had markedly lower baseline test scores than non-adopters. In 2018 (the last pre-policy year), early adopters scored substantial standard deviations below non-adopting districts in STAAR reading and math on average. Late adopters also tended to have somewhat lower pre-TIA scores than non-adopters. However, the gaps were smaller, suggesting that even relatively average-performing districts began to join as the program matured. This is also reflected in the number of districts that adopted the program each year; while 14 districts were the pioneers, 131 districts adopted the program by 2023.

At the district level, early-adopting districts were generally larger in enrollment and more urban/suburban than the typical non-adopter. For instance, the 2019 cohort of adopters had an average district enrollment about 11,000 students higher than non-adopters, this is because some of the state's biggest districts were among the first to implement TIA. Correspondingly, only 46% of those early adopters were classified as rural, compared to 84% of non-adopting districts. Adopting districts also differed on measures of socioeconomic disadvantage. We observe that non-adopting districts had, on average, a *"vulnerability"* index lower than adopters (2.20 for rural non-adopters, 0.85 for non-rural), reflecting significant student vulnerability.[4]

Teachers from adopter and non-adopter districts also differ. Early adopters had slightly younger and less experienced teacher workforces. In 2019 and 2020, the average teacher in adopters ' cohorts had about two fewer years of experience than in non-adopting districts. These gaps diminish in later adoption waves. The share of teachers with a master's or doctorate was similar between adopters and non-adopters. However, early adopters employed a more diverse teaching staff; by 2019, adopters had a 21 percentage point higher share of Hispanic teachers and 10 percentage points more Black teachers than non-adopters. This likely reflects the demographics of large urban districts that joined early. In addition, I do not see large imbalances in terms of core-subject staffing or gender.

In summary, Table 2 makes clear that the selection into TIA was non-random. Adopting districts were typically larger, lower-performing, and served higher-need student populations than those that did not participate. Many small and rural districts initially delayed adoption—either due to administrative capacity constraints or reluctance to alter existing compensation structures—but, over time, TIA gradually expanded into these areas.

These patterns motivate the use of a difference-in-differences research design, which does not require identical baseline characteristics between treated and untreated groups. Instead, the difference-in-differences approach relies on the assumption that, in the absence of adoption, adopters and non-adopters would have followed parallel trends in the outcome. Table 3 examines changes between 2013 and 2018 for both adopting and non-adopting districts. The first column compares non-adopters with early adopters, and the second column compares non-adopters with late adopters. The differences in growth rates across groups are much smaller than the initial level differences and, in most cases, statistically insignificant, providing preliminary empirical support for the parallel trends assumption underlying the identification strategy.

[4] I separate the vulnerability level by rural and non-rural status because, for rural districts, the vulnerability index mechanically gets an increase in the vulnerability score.

Table 2: Descriptive Statistics by Time of Adoption Waves

	Mean	Difference adopters-non adopters by wave	
	Non adopters	2019 to 2021	2022 to 2024
Normalized STAAR (School-grade-level)			
Reading	0.171	−0.398***	−0.203***
		(0.083)	(0.055)
Math	0.138	0.357***	0.162***
		(0.103)	(0.053)
School-level			
Enrollment	538	74.399***	102.557***
		(29.979)	(29.186)
Student to teacher ratio	14.5	1.127***	0.972***
		(0.191)	(0.167)
Nr. schools	2,769	573	2,153
School district-level			
Share of rural	0.84	−0.180***	−0.122***
		(0.071)	(0.029)
Enrollment	3,239	7,801***	2,967***
		(2,668)	(802)
Weeks hybrid in COVID in 2020	5.5	3.577*	3.380***
		(1.854)	(0.819)
Weeks remote in COVID in 2020	0.6	0.888	0.994***
		(0.597)	(0.284)
Vulnerab. score rural districts	2.20	0.331***	0.083
		(0.100)	(0.052)
Vulnerab. score non-rural districts	0.85	1.013***	0.574***
		(0.157)	(0.099)
Nr. districts	636	47	329
Teacher-level			
Experience	11.7	−2.012***	−0.751***
		(0.576)	(0.208)
Age	41.2	−1.546***	−0.799***
		(0.585)	(0.223)
Pay	55,331	-545	127
		(953)	(690)
Share math teachers	0.332	-0.005	-0.008
		(0.014)	(0.012)
Share english teachers	0.374	0.003	0.002
		(0.016)	(0.012)
Share female teachers	0.771	−0.024***	-0.003
		(0.010)	(0.005)
Share hispanic teachers	0.217	0.110*	0.089**
		(0.063)	(0.041)
Share black teachers	0.062	0.070***	0.034**
		(0.022)	(0.017)
Share with master or + degree	0.257	-0.006	-0.014
		(0.012)	(0.010)
Nr. substitute teachers	318	314	382
Nr. teachers	140,658	43,739	137,614

Note: In the first column of results, the table reports the average values of several variables for school districts that have not implemented the policy. The remaining columns display the differences between non-adopters and districts that adopted the policy in different waves (2019-2024). The second column presents the difference between districts adopting the policy between 2019 and 2021 and non-adopters. The last column presents the difference between districts adopting the policy between 2022 and 2024 and non-adopters. Standard errors of the difference between adopters and non-adopters are always clustered at the school district level and are displayed in parentheses. Symbols represent: * p-value<0.1, ** p-value<0.05, and *** p-value<0.01, for tests of differences in mean between non adopters and the adopters for each group. The first panel presents math and reading standardized test scores (STAAR) at the school-by-grade level, the second panel summarizes school-level characteristics, the third panel reports district-level characteristics, and the fourth panel shows teacher-level variables. Standard errors of the average differences are reported in parentheses.

Table 3: Changes Between 2013 and 2018 and Between Adopters and Non-Adopters for Selected Variables by Adoption Waves

	Differences in differences from 2013 to 2018 between adopters-non adopters	
	2019 to 2021 vs non-adopters	2022 to 2024 vs non-adopters
Normalized STAAR (School-grade-level)		
Δ Reading	-0.048	−0.046**
	(0.033)	(0.022)
Δ Math	-0.002	-0.011
	(0.069)	(0.032)
Teacher Characteristics		
Δ Experience	0.111	-0.019
	(0.277)	(0.177)
Δ Age	0.698	-0.269
	(0.569)	(0.270)
Δ Pay	701*	365*
	(391)	(207)
Δ Share math teachers	0.000	-0.003
	(0.012)	(0.007)
Δ Share english teachers	-0.004	0.007
	(0.014)	(0.007)
Δ Share female teachers	−0.015**	0.009*
	(0.007)	(0.005)
Δ Share hispanic teachers	0.011	0.009
	(0.011)	(0.006)
Δ Share black teachers	0.026***	0.009***
	(0.007)	(0.003)
Δ Share with master or + degree	0.015	-0.001
	(0.012)	(0.006)

Note: The columns display the differences between non-adopters and districts that adopted the policy in different waves (2019-2024) in the variation between 2013 and 2018 (i.e., a difference in differences for pre-treatment periods). The first column presents the difference-in-differences between districts that adopted the policy between 2019 and 2021 and non-adopters. The second column presents the difference-in-differences between districts that adopted the policy in 2022 and 2024 and non-adopters. Standard errors of the difference-in-differences are always clustered at the school district level and displayed in parentheses. Symbols represent: * p-value<0.1, ** p-value<0.05, and *** p-value<0.01. The first panel presents math and reading standardized test scores (STAAR) at the school-by-grade level, and the second panel shows teacher-level variables.

4 Impact on Academic Performance

4.1 Empirical Strategy

In this section, I turn the analysis towards evaluating the effect of TIA on performance. I examine TIA's impact on student academic achievement using standardized test scores in math and reading from third to eighth grade. The identification strategy exploits the staggered rollout: essentially, the event-study design compares the change in test scores within a school in a TIA district before vs. after its adoption to the contemporaneous change in schools in districts that have not yet adopted TIA. The following equation can mathematically represent this,

$$Y_{isgt} = \sum_{k \neq -1} \beta_k \cdot \text{TIA}^{(k)}_{D(i)t} + \mathbf{X}'_{it}\gamma + \nu_{R(i)gt} + \lambda_{isg} + \varepsilon_{isgt} \tag{1}$$

where Y_{isgt} represents the standardized test score for school i in district $D(i)$, grade g, subject $s \in \{\text{reading}, \text{math}\}$, and year t. The term $\text{TIA}^{(k)}_{D(i)t}$ denotes a set of event-time indicators that take the value of one if the observation is k years before or after the policy implementation in district $D(i)$, excluding $k = -1$ which serves as the baseline year.

The estimation includes school-by-grade-by-subject fixed effects λ_{isg} to account for baseline differences in this type of staggered program and are documented in section 3.2, and region-by-year effects to flexibly absorb trends $\nu_{R(i)gt}$ at the educational region level. Texas's 20 education regions serve as the primary administrative divisions for implementing state education policies. Each region assists districts with compliance, funding, and teacher evaluation processes, including those related to TIA. The empirical strategy includes region-by-year fixed effects to absorb common shocks and administrative differences across educational regions.

I also include two set of control variables (represented in $\mathbf{X}_{it}$): the interaction between year and quintiles of the number of weeks in hybrid teaching mode, and the interaction between year and quintiles of the number of weeks in remote teaching mode. With these variables, I aim to control for differences in COVID-19 impacts across districts that are closely related to potential divergences in educational outcomes. Finally, ε_{isgt} denotes the error term. Because treatment is assigned at the district level, standard errors are clustered accordingly to allow for arbitrary correlation of unobserved shocks within districts over time.

Recent research has cautioned against the potential pitfalls of implementing staggered event-study designs with traditional TWFE methods. My main estimations used the method proposed by Callaway and Sant'Anna (2021). In particular, the timing of adoption might be correlated with heterogeneous treatment effects. In Appendix B.3, I show a wide range of variations in the specification, including traditional estimation methods, Sant'Anna and Zhao (2020) double robust inverse probability weighting methods, among others, and results remain similar.

4.2 Effect on Test Scores

The dynamic effects illustrated in Figure 3 reports the overall impact on combined math and reading scores. Reassuringly, I do not observe any pre-treatment score trends that would suggest bias due to the violation of parallel trend assumptions. Overall, TIA's impact on learning grows over time, consistent with an initial adjustment period and accumulated benefits over time. [5]

Table 4 summarizes the estimated **group average treatment effects** on the treated on STAAR performance. Column 1 reports the overall impact on combined math and reading scores, while columns 2 and 3 show results for math and reading separately. I find a positive and statistically significant effect of TIA on student achievement. On average, TIA implementation led to an increase of about 0.041 standard deviations (s.d.) in students' test scores, which is statistically significant ad 1% level. These gains are meaningful in the educational context. In subject-specific terms, math scores increased by approximately 0.047 s.d. and reading scores by 0.035 s.d..

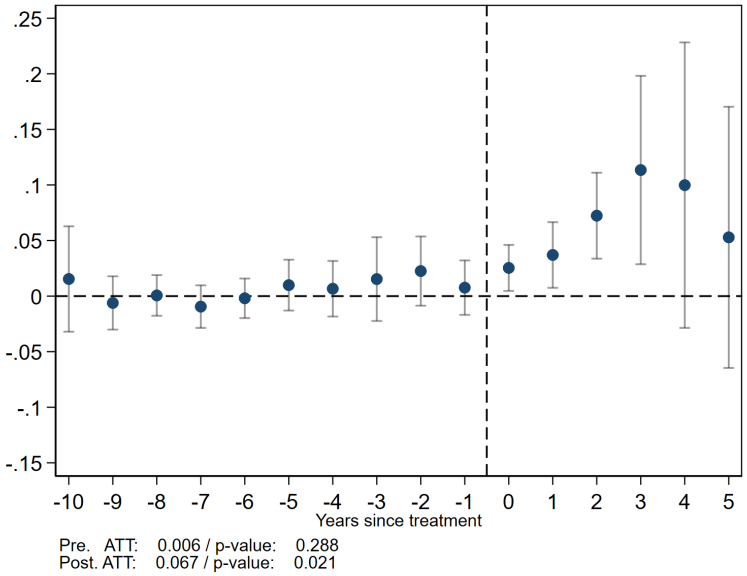
Contextualizing these results relative to prior research, the average magnitude of the TIA impact (0.040 s.d. overall) is slightly below past performance-pay interventions in developing countries, but above what is found in several studies in developed countries. Evidence from developing countries shows that teacher incentives yielded gains in the order of 0.1 s.d. (Muralidharan and Sundararaman, 2011;

[5] Note that the post-treatment average treatment effect is 0.067 s.d., bigger than the group average treatment effect. This happens because the dynamic effects and the group effects weigh cohorts differently.

Glewwe et al., 2010). By contrast, few studies had found essentially zero effect, such as Fryer (2013) experiment in New York City, and Barrera-Osorio and Raju (2017) in Pakistan.

Texas's TIA demonstrates that meaningful improvements in achievement can be achieved and materialized in a short time frame through incentive programs in an extensive decentralized public education system.

Figure 3: Results on Test Scores, Math & Reading Combined (effects in standard deviations)



Note: Dynamic effects calculated using Callaway and Sant'Anna (2021) estimation methods. Outcomes are standardized measures of math and reading scores of the STAAR test, a test implemented in all public schools between the 3rd and 8th grade levels. It includes school-by-subject-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Notice that, in contrast to conventional TWFE, Callaway and Sant'Anna (2021) calculates differences cohort by cohort, comparing scores with respect to not-yet treated observations. This makes it possible to calculate the difference between each pre-treatment period and the not-yet-treated, which means that no pre-treatment coefficient is deleted. Standard errors are clustered at the school district level.

Table 4: Group Average Treatment Effects on STAAR Test Scores

	(1) Math and Reading	(2) Math	(3) Reading
Group ATT	0.0408***	0.0468***	0.0346***
	(3.66)	(2.91)	(3.10)
Region × year F.E.	✓	✓	✓
COVID Controls *times* Year F.E. × year F.E.	✓	✓	✓
School × Subject × Grade level F.E.	✓	✗	✗
School × Grade level F.E.	✗	✓	✓
Observations	321,132	160,116	161,016
Nr. Untreated schools	2,250	2,250	2,248
Nr. Treated schools	2,469	2,467	2,468
Outcome Av. ctrl	0.029	0.014	0.044

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Group Average Treatment Effect calculated using Callaway and Sant'Anna (2021) estimation methods. Outcomes are standardized measures of math and reading scores of the STAAR test, a test implemented in all public schools between the 3rd and 8th grade levels from 2013 to 2024. Observations used in this regression are at the school year-by-school-by-subject-by-grade level, and I weight by the number of test takers. I use data from 2010 to 2023. It includes school-by-subject-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Notice that the Group Average Treatment Effect in Callaway and Sant'Anna (2021) weights observations differently from the conventional TWFE methodology. Therefore, coefficients might differ; I checked that, in this case, results do not change substantially. Standard errors are clustered at the school district level.

4.3 Robustness Checks

I provide a range of robustness checks to address concerns over specific confounders or decisions regarding the estimation method. Applying multiple estimation approaches yields consistent results across specifications. In addition to the traditional two-way fixed effects (TWFE) model, I estimate the effects using the double-robust inverse probability weighting method proposed by Sant'Anna and Zhao (2020). As shown in Appendix B.3, both approaches produce similar magnitudes and patterns of estimated effects, reinforcing the robustness of the main findings. The specification that use only adopting districts- and therefore uses variation in adoption timing for identification, shows less precise coefficients, but there is still a positive effect during the first years after adopting TIA.

A potential concern is that the onset of COVID-19 coincided with the implementation of TIA in several early-adopting districts. If pandemic severity is correlated with program adoption and also affects student outcomes, the estimates could be biased. The main specification addresses this concern by including COVID-related variables interacted with year fixed effects, capturing differential pandemic effects across time and place.

As an additional robustness check, Appendix B.3 reports estimates that exclude districts adopting between 2019 and 2021—the period most affected by COVID disruptions. This specification identifies the policy's effect only for the first three years after implementation, and the resulting coefficients for those years are nearly identical to those obtained in the baseline specification. I also estimate versions that omit the COVID-related control variables entirely; these yield similar coefficients, suggesting that pandemic-related dynamics are not a major threat to identification.

Finally, I estimate a model without region-by-year fixed effects. The results remain highly consistent, with slightly attenuated magnitudes but unchanged overall patterns, confirming that the findings are

not driven by regional time-varying shocks.

A second potential concern is that the estimated effects could be driven by heterogeneous trends across districts adopting the program in different years. Table B2 (and Appendix B.8) reports average treatment effects by cohort of adoption. The results align with expectations for each adoption wave: test score gains appear most clearly among earlier cohorts. This pattern reflect the longer exposure under the program for early adopters, who have had more time for the policy to take effect. For example, districts that adopted in 2022—exposed to the program for three academic years by 2024—show significant test score gains of approximately 0.05 standard deviations. Earlier adopters (the 2019 and 2020 cohorts) eventually experienced improvements ranging from 0.036 to 0.093 standard deviations, although these estimates are less precise and in some cases not statistically significant at conventional levels. In contrast, the most recent adopters (2023 and 2024 cohorts) show no measurable effects yet, as insufficient time has elapsed for the policy to influence outcomes.

In Appendix B.5, Figure B3, I show the coefficient of the average treatment effect dropping one treated school district at a time. In all the specifications, coefficients are statistically different than zero with an average treatment effect close to 0.04 s.d., showing that outliers do not drive results.

In the following sections, I explore the mechanisms underlying these gains, focusing on how changes in the teacher workforce may have contributed to the observed improvements.

4.4 Effect on Teachers' Composition

A performance pay program can influence student achievement via several channels. One channel is incentivizing greater teacher effort and effectiveness. Another channel could be the change in the composition of the workforce. By offering substantial financial rewards, the program could retain effective teachers who might otherwise leave, particularly in high-need schools (Podgursky and Springer, 2007). Finally, the program could reduce attrition; if teachers decide to stay instead of leaving the workforce or moving to another school, students will benefit from the teachers' familiarity with the classroom, increasing classroom stability, and even avoiding schools' costs associated with teacher shortages.

In this section, I examine teacher workforce composition, asking whether schools in TIA districts saw changes in the experience level, tenure, teachers' credentials, teachers' exits, and entries.

Average experience and the share of teachers with graduate education are commonly used in the literature as proxies of quality. To assess whether these variables serve as credible proxies for teacher quality, I examine the characteristics associated with the likelihood of receiving a designation (Appendix B.1). Teachers' education, experience, and tenure are positively correlated with the probability of receiving a designation, although only experience and tenure are statistically significant at conventional levels.

Similarly to the previous section, the identification strategy is represented in the following equation:

$$Y_{it} = \sum_{k \neq -1} \beta_k \cdot \text{TIA}^{(k)}_{D(i)t} + \mathbf{X}'_{it}\gamma + \nu_{R(i)t} + \lambda_i + \varepsilon_{it} \tag{2}$$

Now each observation represents one school i at a particular year t. The outcomes are the average year

of experience of teachers in the school, the share of teachers in the school with post-graduate degrees,[6] the percentage of teachers in the school who leave the next year, and the percentage of teachers who are new. I call these variables *"school exit"* and *"school entry"*.[7] In all specifications, I include the same COVID-related control variables as in the previous specification. I also add educational region-by-year fixed effects $\nu_{R(i)t}$, and school fixed effects λ_i. Standard errors are clustered at the school-district level.

In the first few years after adoption, the average teacher tenure and experience in treated schools rose relative to control schools. Figure 4a plots the event-study estimates for mean years of experience: the trajectory for TIA schools increases from years two to three after the implementation of the policy. This resonates with the improvement in retention I document (discussed in the following paragraphs). The share of teachers with advanced degrees (master's or higher) in TIA schools increase slightly after two years into the program, however, this increase is not statistically significant at conventional levels. (Figure 4c).

The policy also appears to affect teacher mobility, although these effects are less precisely estimated. Figures 4e and 4d plot teacher exit and entry rates. Teacher attrition falls in the second year after adoption, and teacher entries decline beginning in the third year after adoption. In both cases, the effects subsequently dissipate and return to approximately zero, albeit with wide confidence intervals. This pattern is consistent with a temporary pause in turnover between the second and third year following implementation. A plausible interpretation is that teachers respond strategically to the prospect of receiving a designation. After the policy is introduced, teachers first accumulate a baseline year of performance and then remain in place for at least one additional year so that they can be evaluated based on student growth. Once designations are announced, teachers decide whether to remain or exit under incentives that more closely resemble pre-policy conditions.

These results align with the observed rise in average teacher experience and tenure. When teachers choose to stay an additional year, their average experience naturally increases. This effect is amplified if the reduction in exits occurs primarily among already experienced teachers. Indeed, the heterogeneity analysis by experience level shows that the decline in teacher exits is concentrated among experienced staff (Figures B4 and B5 in the appendix). Thus, the increase in average experience is largely explained by greater retention of experienced teachers in TIA districts.

Schools participating in TIA exhibit lower attrition among experienced teachers, gradually raising the average tenure of their teaching force. Still, the overall changes in teacher mobility are modest. Such limited short-term effects are expected: earning a TIA designation requires meeting sustained effectiveness benchmarks, which may make the bonuses less immediately attainable for newly hired teachers. Moreover, because only a small share of teachers receive a designation, the aggregate reduction in exits is driven mostly by experienced teachers—those most likely to qualify for the program (as shown in Section B.1).

In short, TIA did not dramatically alter the composition of the teacher's educational background, although the results suggest an increase in experience. This indicates that test score gains were not driven

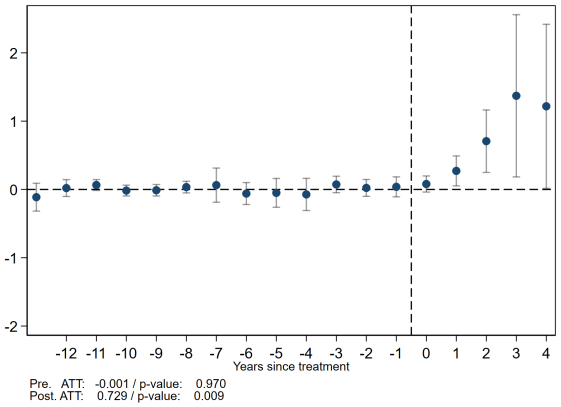
[6] This includes teachers with master's and doctoral degrees, although mostly master's degrees.

[7] Formally, school exit is $\frac{\text{Left school}_{i,t+1}}{\text{Nr. teachers}_{i,t}}$ and school entry is $\frac{\text{Arrive to district}_{i,t}}{\text{Nr. teachers}_{i,t-1}}$.

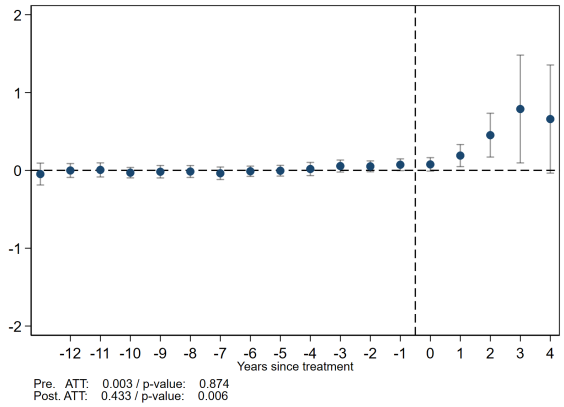
by an influx of highly educated teachers or the exit of lower-educated ones. My results reinforce that financial incentives can be an effective tool to retain teachers. By reducing turnover, TIA may indirectly raise instructional quality, since teacher effectiveness generally improves with experience, especially in the early career years (Papay and Kraft, 2015). Keeping effective teachers on the job and in the same schools is a plausible mechanism explaining the gains observed in standardized test scores. In the next section, I will explore whether the gains in test scores attributed to the program are explained by these changes in teachers' workforce composition, or other plausible mechanisms, such as increases in effort or changes in managerial practices.

Figure 4: Results on Schools' Teachers Composition and Movement (effects in percentage points)

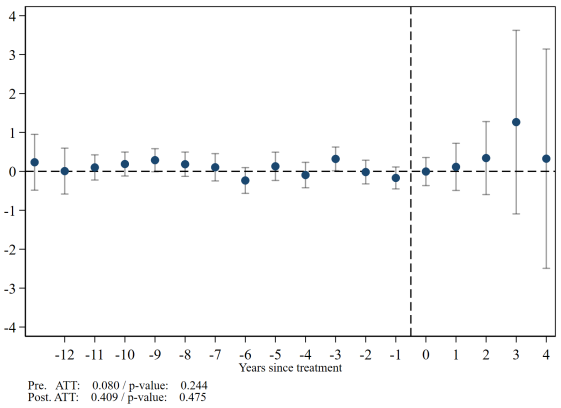
(a) Average Years of Experience



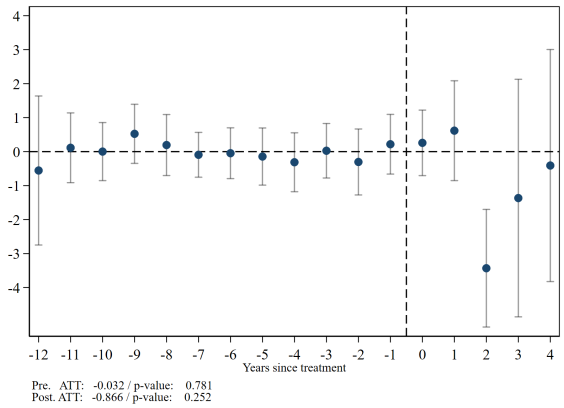
(b) Average Years of Tenure



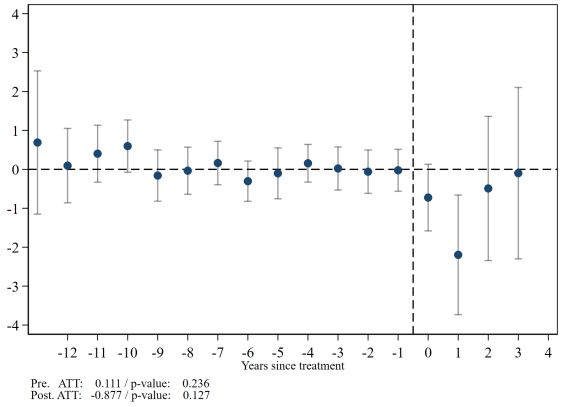
(c) % Teachers with Post-graduate Studies



(d) School Entry Rate (percentage)



(e) School Exit Rate (percentage)



Note: Dynamic effects calculated using Callaway and Sant'Anna (2021) estimation methods. Outcomes are: the average year of experience of each school, the average years of the teacher teaching in that school (tenure), the share of teachers in the school that has postgraduate studies (mostly masters' degrees), the share of teachers who are new in the school (School Entries), and the number of teachers that are going to leave the school next year (School Exits, using data up to 2022 because of data constraints). The estimation includes school-level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Notice that, in contrast to conventional TWFE, Callaway and Sant'Anna (2021) calculates differences cohort by cohort, comparing scores with respect to not-yet treated observations. This makes it possible to calculate the difference between each pre-treatment period and the not-yet-treated, which means that no pre-treatment coefficient is deleted. Standard errors are clustered at the school district level.

Table 5: Average Treatment Effect on Schools' Teachers Composition and Movements (effects in percentage points)

	(1) Av. Experience	(2) Av. Tenure	(3) Share Postgr.	(4) Turnover	(5) Entry rate	(6) Exit rate
TIA	0.146	0.158**	0.168	-0.996**	-1.156**	-0.700*
	(0.100)	(0.0682)	(0.325)	(0.487)	(0.542)	(0.411)
Ctrls. × year	✓	✓	✓	✓	✓	✓
School × subject × grade level F.E.	✗	✗	✗	✗	✗	✗
School F.E.	✓	✓	✓	✓	✓	✓
District × year F.E.	✗	✗	✗	✗	✗	✗
Observations	48,131	48,131	43,846	39,384	43,846	43,846
Nr. Untreated schools	2,192	2,192	2,190	2,188	2,190	2,190
Nr. Treated schools	2,444	2,443	2,444	2,443	2,444	2,444
Outcome Av. ctrl	9.865	6.145	21.612	16.725	16.717	15.573

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Average Treatment Effect calculated using conventional TWFE methodology. Outcomes are: the average year of experience of each school, the average years of the teacher teaching in that school (tenure), the share of teachers in the school that has postgraduate studies (mostly masters' degrees), the share of teachers who are new in the school (School Entries), and the number of teachers that are going to leave the school next year (School Exits). Observations used in this regression are at the school year-by-school level, and I weight by the number of test takers, I use data from 2013 to 2023. It includes school-by-subject-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

4.5 Is the Impact Driven by Changes in Teacher Composition?

A key question is whether the observed improvements in student achievement following the implementation of the Teacher Incentive Allotment (TIA) reflect changes in the composition of the teaching workforce. Because the program rewards high-performing teachers, it may encourage experienced and effective teachers to remain in the classroom, thereby gradually shifting the average quality of teachers in participating schools. This section assesses how much of the program's total impact can be attributed to such compositional shifts (a detailed explanation of the process is available in appendix B.8).

In theory, we could disentangle this relationship using instrumental variables to isolate the exogenous variation in experience. In practice, however, finding instruments that shape teacher retention without simultaneously influencing student learning has proven particularly challenging.[8] To isolate the portion of achievement explained by teachers' observable characteristics, I employ two complementary strategies. The first constructs a teacher-composition index that captures the share of test score variation driven by measurable attributes such as age, tenure, experience, education, and race, interacted with subject and grade levels. Using pre-policy data (2010–2018), I estimate the relationship between these attributes and test performance, applying LASSO to select the most predictive features. The fitted values from this model represent the *"teacher-explained"* component of achievement. Estimating the average treatment effect on this component allows me to assess how much of the TIA effect operates through observable changes in the composition of teachers.

The second approach uses a mediation framework to decompose the total policy effect on test scores into direct and indirect components. I include potential mediators, such as average experience, retention, turnover, and the predicted *teacher-explained score* measure, along with flexible interactions by subject and grade. This specification identifies the portion of the total effect that can be statistically attributed

[8] For instance, using changes on teachers' outside options could indirectly influence the local economy of students' parents, thereby violating the exclusion restriction. Moreover, other shift–share instruments used in turnover analyses often suffer from regression-to-the-mean concerns or limited statistical power.

to workforce shifts, under the assumption that no omitted factors simultaneously affect both mediators and outcomes.

Results of the first approach are reported in Table 6. The treatment effect on the *teacher-explained* component of achievement accounts for roughly 6% (0.00187/0.0316) of the overall effect. However, the impact on the *teacher-explained* component is not statistically different from zero. The mediation analysis (Table 7) yields similar results: depending on the specification, the share of the policy effect mediated by teacher composition and retention ranges between 2 and 14 %, with the upper bound obtained when incorporating detailed controls for teacher inflows, outflows, and demographics. In all cases, the majority of the effect remains unexplained by measurable workforce changes.

These findings suggest that the main driver of student achievement under TIA lies beyond observable compositional shifts. The evidence instead points to behavioral and organizational mechanisms as the most plausible channels, such as increased teacher effort, improved instructional practices, or stronger school management. Reassuringly, the modest contribution of workforce composition is slightly smaller in magnitude than the one-fifth share attributed to teacher selection in **?**, which implements an experiment to exclusively isolate for compositional effects.

Table 6: Average Treatment Effect on Test Scores
and Test Score Predicted by Changes in Schools' Teachers Composition

	(1) Test Scores	(2) $\hat{Y}(teachers)$
TIA	0.0316**	0.00187
	(0.0148)	(0.00243)
Ctrls. × year	✓	✓
School × subject × grade level F.E.	✓	✓
School F.E.	✗	✗
District × year F.E.	✗	✗
Observations	263,140	262,614
Nr. Untreated schools	2,202	2,196
Nr. Treated schools	2,454	2,446
Outcome Av. ctrl	0.145	0.020

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Average Treatment Effect calculated using conventional TWFE methodology. Outcomes are standardized measures of math and reading scores of the STAAR test, and the STAAR test score predicted by teachers composition in the school. Observations used in this regression are at the school year-by-school-by-subject-by-grade level, and I weight by the number of test takers, I use data from 2013 to 2023. It includes school-by-subject-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

Table 7: Average Treatment Effect Including Changes in Composition and Retention as Control

	(1)	(2)	(3)	(4)	(5)
TIA	0.0316**	0.0309**	0.0305**	0.0300**	0.0269**
	(0.0148)	(0.0146)	(0.0144)	(0.0143)	(0.0135)
Mediators	None	Experience	$\hat{Y}(teachers)$	Experience + $\hat{Y}(teachers)$	Experience + $\hat{Y}(teachers)$+ others
Observations	263,140	263,140	262,614	262,614	261,917
Nr. Untreated schools	4,656	4,656	4,642	4,642	4,642
Nr. Treated schools	2,007	2,007	1,999	1,999	1,999
Outcome Av. ctrl	0.145	0.145	0.145	0.145	0.145

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Each column shows different measures of average treatment effects, including different variables as control. Outcomes are standardized measures of math and reading scores of the STAAR test. In this table, the results show coefficients calculated with conventional TWFE, with data from 2013 to 2023. All specifications include interactions of COVID-related variables with year, grade-by-subject-by-year fixed effects, and region-by-year fixed effects. The row "Mediation" indicates which other control variables are included. Experience stands for the average years of experience of math and English teachers in the school, and the score is projected with a correlation between the score and pre-2019 characteristics ($\hat{Y}(teachers)$). The last column includes several control variables: school's average age, school's average experience, $\hat{Y}(teachers)$, school's number of female teachers, and school's number of teachers by race. All these variables interacted with subject and grade level. Standard errors are clustered at the school district level.

5 Closing the Gap: Effects on Test Scores by School Vulnerability

By design, the size of the allotment is positively related to school vulnerability. This design feature not only reflects the program's redistributive intent but also raises an empirical question: are more vulnerable schools having larger improvements in student outcomes? If the effect of the policy is stronger in high-vulnerability schools, this would indicate that resources are being directed where they can have the greatest impact, enhancing both students' achievement and equality in education. Moreover, analyzing differential impacts across vulnerability levels provides insight into whether the program is helping to close the educational performance gap. Vulnerable schools typically start at a lower baseline achievement level, face greater staffing challenges, and have lower implementation capacities. Detecting larger effects in these schools would therefore suggest that the policy contributes to convergence in educational outcomes, rather than widening disparities. In this section, I analyze the impact of the policy on schools with different levels of vulnerabilities. Then I also examine intra-district inequality to see if these heterogeneous effects persist among schools under the same local government, reducing concerns about endogeneity issues that affect adoption timing.

5.1 Heterogeneous Effect on Test Scores by School Vulnerability

To measure the heterogeneous effect of the TIA at different levels of vulnerability, I implement an identification strategy represented by the following equation:

$$Y_{isgt} = \alpha_0 + \alpha_1 TIA_{D(i),t} + \alpha_2 TIA_{D(i),t} \times Vulnerability_i + \mathbf{X}'_{it}\gamma + \nu_{R(i)t} + \lambda_{isg} + \delta_t + \varepsilon_{isgt} \quad (3)$$

where Y_{isgt} is either the standardized test score of school i, in subject s, grade-level g, and year t, or the estimated score explained by characteristics of the teachers in the schools. Additional analyses involve a similar regression, but using outcomes of schools' teachers' composition at, in this case, observations are at the school-by-year level, and instead of using school-by-subject-by-grade level fixed effects (λ_{isg}), I include school-level fixed effects. Standard errors are clustered at the school-district level.

The evidence indicates a clear positive relationship between school vulnerability and test score gains. More vulnerable schools, which receive larger incentive allotments per student, experienced greater improvements in test scores. Figure 5 illustrates this pattern: it shows that schools with higher vulnerability indices registered higher test score increments after the program. This visual pattern is confirmed by the regression estimates in Table 8, which interact the treatment effect with a continuous measure of school vulnerability linearly as specified in equation 3. This vulnerability measure is a school-level index calculated by the *Texas Educational Agency* that goes from 0 (least vulnerable) to 4 (most vulnerable), and it is based on the socioeconomic characteristics of the students and school rurality. The coefficient TIA is negative, suggesting a negative effect in the richest places. Still, the interaction term is positive and significant, indicating that the test score improvement attributable to the program grows with the level of school vulnerability and becomes positive after a vulnerability index value of 1 (which are still relatively rich schools). In practical terms, a one-point increase in the vulnerability index is associated with a considerably large program effect on test scores.

These results suggest that the program effectively targeted and benefited the schools most in need. However, it is important to consider whether part of the observed improvements may stem from compositional changes in the teacher workforce rather than from shifts in teaching practices or effort. One might wonder, for instance, whether the substantial test score gains in vulnerable schools reflect changes in teachers' observable characteristics, such as experience or tenure, rather than direct productivity improvements.

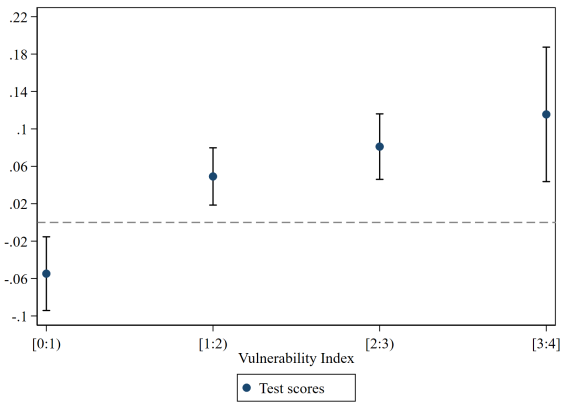
Earlier, I showed that teacher composition plays a limited role in explaining the average effect. Yet, the mechanisms operating in schools with different vulnerability levels may differ, motivating a closer examination of how changes in teacher composition contribute to test score gains across the vulnerability distribution. The evidence indicates that compositional factors alone cannot account for the magnitude of the improvements. Teacher characteristics evolve in a more favorable direction for vulnerable schools: Table 8 shows that average experience and tenure increase more in these schools, while turnover and teacher mobility decline more sharply. Similarly, the component of test score variation attributable to changes in teacher composition improves more in vulnerable schools (Table 8, second column). Nevertheless, the size of this effect remains modest. A back-of-the-envelope calculation suggests

that changes in teacher characteristics explain only about 5% of the total test score impact across most vulnerability levels, indicating that most of the gains likely arise from other channels —such as increased teacher effort or improved school management practices.

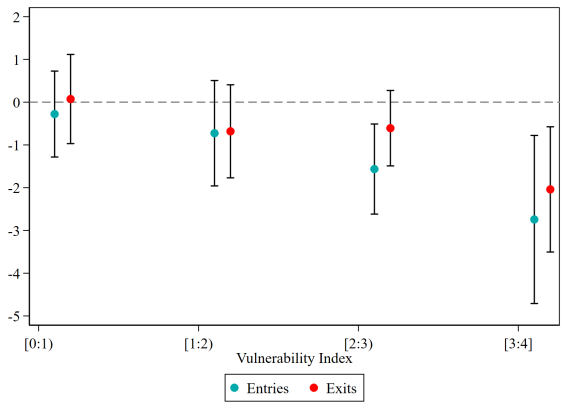
It is worth noting an important identification challenge in interpreting these results: school vulnerability is inherently correlated with the size of the incentive allotment. By program design, the more vulnerable schools received larger grants, so vulnerability and treatment intensity go hand-in-hand. This correlation makes it difficult to disentangle whether the greater test score gains in vulnerable schools were due to their vulnerability per se (and related unobserved factors) or simply the effect of receiving more resources. These findings, however, already suggest that the program's design of channeling greater support to higher-need schools either resulted in markedly larger test score improvements for those schools or shows that resources are allocated to schools that are able to achieve larger improvements.

Figure 5: Results by Ranges of Vulnerability

(a) STAAR Scores (s.d.)



(b) Exits and Entries (perc. points)



Note: Average Treatment Effect by vulnerability level calculated using conventional TWFE methodology. Observations used in the first figure are at the school year-by-school-by-subject-by-grade level, and I weight by the number of test takers. In the second figure, observations are are the year-by-school level. I use data from 2010 to 2023. It includes school-by-subject-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

Table 8: Heterogeneous Treatment Effect on Test Scores and School-level Changes in Teachers' Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores	$\hat{Y}(teachers)$	Av. Experience	Av. Tenure	Share Postgr.	Turnover	Entry rate	Exit rate
TIA	-0.0712***	-0.00260	-0.197	0.0104	0.00847	0.663	0.183	0.320
	(0.0222)	(0.00293)	(0.126)	(0.0962)	(0.620)	(0.584)	(0.550)	(0.600)
TIA × Vulnerability	0.0634***	0.00276**	0.190***	0.0815*	0.0843	-0.881***	-0.746***	-0.541**
	(0.0116)	(0.00140)	(0.0610)	(0.0424)	(0.297)	(0.264)	(0.279)	(0.252)
Ctrls. × year	✓	✓	✓	✓	✓	✓	✓	✓
School × subject × grade level F.E.	✓	✓	✗	✗	✗	✗	✗	✗
School F.E.	✗	✗	✓	✓	✓	✓	✓	✓
District × year F.E.	✗	✗	✗	✗	✗	✗	✗	✗
Observations	263,080	262,554	48,121	48,016	43,837	39,377	43,662	43,837
Nr. Untreated schools	2,201	2,195	2,191	2,191	2,189	2,187	2,189	2,189
Nr. Treated schools	2,454	2,446	2,444	2,443	2,444	2,443	2,444	2,444
Outcome Av. ctrl	0.145	0.020	9.866	6.145	21.618	16.695	16.697	15.557

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Treatment Effect by vulnerability calculated using conventional TWFE methodology. Observations used in the first two columns of the results are at the year-by-school-by-subject-by-grade level, and I weight by the number of test takers. In these two colums, the outcomes are standardized measures of math and reading scores at the school-by-subject-by-grade-level, and the predicted score by using teachers composition of the school. The rest is at year-by-school level. I use data from 2010 to 2023. It includes school-by-subject-by-grade level fixed effects or year-by-school when applicable, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Av. Experience is the average years of experience of teachers in school i, Av. Tenure is the average years of teachers in working in school i, Share Postgr. is the share of teachers in the school with postgraduate degrees. Turnover is the average of Entry Rate and Exit Rate. Entry Rate is the number of new teachers as a portion of the number of total teachers in the school last period. Exit Rate is the number of teachers that are going to leave next year as a portion of the number of total teachers in the school in the current period. Standard errors are clustered at the school district level.

5.2 Effect on Within-District Test Score Inequality

An important question is how the program affected achievement gaps within districts and year. Cross-district comparisons reflect differences not only in school vulnerability but also in broader district-level characteristics such as changes in management quality, funding structures, or administrative capacity. By contrast, focusing on within-district-year variation control for these district-level changes, allowing for a cleaner comparison between more and less vulnerable schools operating in the same local environment.

Because the policy is implemented at the district level, it is not possible to estimate an average treatment effect within district–year cells, as there is no variation in treatment status across schools within the same district in a given year. Nonetheless, it is still possible to explore heterogeneity in effects across schools within districts over time. I do so using two complementary approaches.

First, I implement an event-study specification similar to equation 3, but instead of including district-level controls, I add district-by-year fixed effects. This specification absorbs any shocks that affect all schools within a district equally and, therefore, identifies differences in policy effects across schools within the same district–year. However, it does not allow for the estimation of a baseline group by vulnerability, as the timing of program adoption is collinear with the district-by-year fixed effects.

Second, I estimate the average effect of the policy on various measures of inequality, calculated at the district-by-subject level. These regressions include district-by-subject fixed effects, region-by-year fixed effects, and year interactions with COVID-related controls. This approach captures how the policy affects disparities across schools within districts using conventional measures of inequality.

Table 9 presents a heterogeneity analysis with district-by-year fixed effects. The table reveals a statistically significant gap in program impact between these schools. Within a given district, a high-vulnerability school saw a much larger increase in test scores post-treatment than a low-vulnerability

school did. Table 9 also report heterogeneous analysis for outcomes related to teacher composition, and shows that this pattern holds consistent across outcomes, although it is not significance at conventional significant levels in all outcomes.

Table 10 turns to district-level measures of test score dispersion to evaluate changes in inequality after the program's implementation. I measure intra-district inequality using metrics such as the Gini coefficient, Atkinson index, the Quantile Coefficient of Dispersion, and the gap between average test scores at the school-grade level within each district and subject (e.g., the gap between the 75th and 25th percentile of school-by-grade level average scores). The table documents a decline in these inequality measures in the post-treatment period in treated districts relative to the pre-treatment baseline. For instance, the within-district Gini coefficient in test scores dropped by roughly 4.6%, indicating that schools' performance levels converged within districts. Likewise, the percentile gap measures shrank between 2.3 and 8.7%. Interestingly, the percentile gaps that reduced the gap the most were those that considered lower percentiles, suggesting that the program improved the scores of school-by-grade-level observations with low average scores.

The fact that test score improvements were concentrated in vulnerable schools has important implications for educational inequality. By lifting the bottom of the distribution substantially, the program effectively reduced achievement disparities among schools. Notably, there is little evidence that the program came at the expense of top-performing schools; their scores generally continued on trend, except for some evidence of adverse effects at the bottom levels of vulnerability, while the lower-performing schools improved at a much faster rate. This suggests an equalizing impact rather than a trade-off. This outcome is encouraging from a policy standpoint: it suggests that interventions of this sort can mitigate entrenched inequalities by bringing up the performance of schools in more challenging environments.

Table 9: Within-district Heterogeneous Treatment Effect on Test Scores and School-level Changes in Teachers' Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores	$\hat{Y}(teachers)$	Av. Experience	Av. Tenure	Share Postgr.	Turnover	Entry rate	Exit rate
TIA × Vulnerability	0.0809***	0.00206***	0.140***	0.0676	0.120	-0.385	-0.240	-0.249
	(0.0149)	(0.000585)	(0.0493)	(0.0415)	(0.337)	(0.263)	(0.257)	(0.244)
Ctrls. × year	✓	✓	✓	✓	✓	✓	✓	✓
School × subject × grade level F.E.	✗	✗	✗	✗	✗	✗	✗	✗
School F.E.	✗	✗	✗	✗	✗	✗	✗	✗
District × year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Observations	263,080	266,924	45,747	45,584	41,685	37,436	41,501	41,685
Nr. Untreated schools	2,201	2,242	2,040	2,040	2,038	2,036	2,040	2,038
Nr. Treated schools	2,454	2,461	2,400	2,400	2,400	2,397	2,398	2,400
Outcome Av. ctrl	0.145	0.016	9.884	6.219	22.125	16.183	16.124	15.166

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Average Treatment Effect calculated using conventional TWFE methodology. Observations used in the first two columns are at the school year-by-school-by-subject-by-grade level, and I weight by the number of test takers. The remaining columns show results at the year-by-school level (which is why the table reports fewer observations). I use data from 2013 to 2023, except for results in exit, entries, and turnover, where one year of data is dropped because it is not possible to compute entries for the first year, exits for the last year, and turnover (average between entries and exits) for the first and last year. It includes school-by-subject-by-grade-level fixed effects, school fixed effects when applicable, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

Table 10: Effect on Measures of Score Gaps at the School District-level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Gini	Atkinson (1)	Quant. Coeff. Disp.	P25–P50	P75–P50	P75–P25	P10–P50	P90–P50	P90–P10
TIA	-0.127***	-0.0102***	-0.257***	-5.211***	-3.260**	-8.471***	-3.840**	-1.614	-5.454**
	(0.0413)	(0.00351)	(0.0653)	(1.515)	(1.452)	(2.095)	(1.903)	(1.862)	(2.632)
Effect %	-4.6 %	-7.4 %	-6.6 %	-8.7 %	-5.2 %	-6.9 %	-2.9 %	-1.6 %	-2.3 %
Observations	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000
Nr. Untreated districts	427	427	427	427	427	427	427	427	427
Nr. Treated districts	323	323	323	323	323	323	323	323	323
Outcome Av. ctrl	2.739	0.138	3.881	60.185	63.205	123.390	130.632	102.964	233.596

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Average Treatment Effect calculated using conventional TWFE methodology. The observations used in this regression are at the school-year-by-school-district-by-subject level. The outcomes, from the first to the last column, are: the Gini coefficient, Atkinson index, Quantile Coefficient of Dispersion, Gap between percentile 50 and 25, Gap between percentile 75 and 50, Gap between percentile 75 and 25, Gap between percentile 50 and 10, Gap between percentile 90 and 50, Gap between percentile 90 and 10, all of them for each school district and subject. I use data from 2013 to 2024. It includes school district-by-subject fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. I exclude small school districts that contain only one school implementing the STAAR test. Standard errors are clustered at the school district level.

5.3 Is the Larger Effect in Vulnerable Schools Driven by Differences in Marginal Productivity or Higher Bonus Allotments?

The positive impact of TIA on student achievement is stronger in more vulnerable schools. Two mechanisms likely explain this pattern. First, these schools may exhibit higher returns to the reform because they begin from lower performance baselines —i.e. improvements in teaching effort or retention translate into larger learning gains. Second, the TIA bonus schedule provides stronger monetary incentives in high-vulnerability settings, directly motivating greater effort and retention.

To distinguish between these channels, I control for school-level factors correlated with vulnerability—such as socioeconomic composition, baseline scores, and student retention—that proxy for differences in productivity. Results in Table 11 show that schools with lower baseline achievement or higher poverty experience larger effects, consistent with higher marginal returns. Yet, even after controlling for these variables, the interaction between TIA and vulnerability remains positive and significant, suggesting that differential financial incentives also play an independent role.

In the most demanding specification, which includes extensive interactions capturing schools' baseline productivity, the TIA–vulnerability interaction continues to be robust. In this specification: a one-point increase in the vulnerability score -roughtly the difference between schools in percentile 50 and 75 of the vulnerability score- corresponds to a 0.07 standard deviation improvement in test scores. Overall, both mechanisms—higher marginal returns and stronger incentives— seems to contribute to the greater policy impact observed in vulnerable schools.

Table 11: Heterogeneous Treatment Effect on Test Scores Including Controls Related with Productivity at School

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TIA	-0.0712***	-0.0574***	-0.0312	-0.0330					
	(0.0222)	(0.0212)	(0.0214)	(0.0213)					
TIA × Vulnerability	0.0634***	0.0567***	0.0367***	0.0378***	0.0809***	0.0753***	0.0468***	0.0466***	0.0795***
	(0.0116)	(0.0111)	(0.0128)	(0.0126)	(0.0149)	(0.0134)	(0.0146)	(0.0146)	(0.0145)
TIA × score 2018 (normalized)		-0.0200**	-0.0186**	-0.0215***		-0.0132	-0.0134	-0.0138	
		(0.00790)	(0.00804)	(0.00818)		(0.00830)	(0.00838)	(0.00882)	
TIA × income/poverty 2014-18 (normalized)			-0.0341**	-0.0338**			-0.0460***	-0.0460***	
			(0.0143)	(0.0142)			(0.0160)	(0.0160)	
TIA × Attrition 2014-18 (normalized)				-0.0254***				-0.00409	
				(0.00650)				(0.00695)	
Ctrls. × year	✓	✓	✓	✓	✗	✗	✗	✗	✗
School × subject × grade level F.E.	✓	✓	✓	✓	✓	✓	✓	✓	✓
District × year F.E.	✗	✗	✗	✗	✓	✓	✓	✓	✓
TIA × year F.E. × 20 quantiles of (score 2018 × income/poverty × Attrition)	✗	✗	✗	✗	✗	✗	✗	✗	✓
Observations	263,080	263,080	263,080	262,040	267,550	267,550	267,550	266,130	259,593
Nr. Untreated schools	2,201	2,201	2,201	2,200	2,249	2,249	2,249	2,247	2,247
Nr. Treated schools	2,454	2,454	2,454	2,453	2,469	2,469	2,469	2,468	2,468
Outcome Av. ctrl	0.145	0.145	0.145	0.144	0.144	0.144	0.144	0.146	0.139

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Average Treatment Effect calculated using conventional TWFE methodology. The outcome is the score on standardized STAAR test. Observations used in this regression are at the school year-by-school district level. I use data from 2010 to 2023. The first four columns include school-by-subject-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Columns four to eight include school-by-subject-by-grade level fixed effects, and school district by year fixed effects. The last column includes includes school-by-subject-by-grade level fixed effects, and school district by year fixed effects, and the interaction between year dummies, TIA dummy, twenty dummies indicating in which percentile the school is in terms of test scores in 2018,twenty dummies indicating in which percentile the neighborhood where the school locate is in terms of income/poverty, and twenty dummies indicating in which percentile the school is in terms of school attrition between 2014 and 2018. Standard errors are clustered at the school district level.

5.4 Cost-Effectiveness Analysis by School Vulnerability

The policy successfully reduces inequality and improves test scores, but it also channels substantial funds to high-performing teachers. This raises the question of whether the program is cost-efficient. While the policy entails several types of costs, I focus on the state's allotments, which represent the largest share of total expenditures. Using this measure, I provide an approximate assessment of the policy's cost-effectiveness. Specifically, I estimate the effect of the policy during its first four years across eight groups of schools, classified by their vulnerability score, and compute the total allotments received by teachers in those schools over the same period. The results are presented in Appendix B.10.

Figure B9 summarizes the relationship between program effects and costs across vulnerability levels during the first four years of TIA implementation. Panel (a) shows the estimated impact of the Teacher Incentive Allotment on student achievement, while Panel (b) reports the average cost per student, calculated from the total allotments received by schools within each vulnerability bin. Program costs per student rise with vulnerability, from roughly $130 per student in the least vulnerable schools to over $400 per student in the most vulnerable. This pattern reflects the policy's design, which deliberately allocates larger bonuses to teachers in disadvantaged schools to attract and retain high-quality educators.

Although per-student costs increase with vulnerability, gains in test scores rise proportionally more. Panel (c) combines effects and costs to compute the implied cost per one–standard deviation increase in STAAR scores per student over the first four years of implementation. The ratio is negative or near zero for the least vulnerable schools, which show small or negative average treatment effects, but becomes positive for mid- and high-vulnerability localities. These results indicate that TIA is most cost-effective in

the schools serving students with the fewest resources. Moreover, the policy compares favorably to other education spending programs. For instance, Jackson and Mackevicius (2024) estimate that increasing per-pupil spending by $32,258 over four years is required to raise test scores by one standard deviation, whereas the most costly group of TIA schools achieves comparable gains at approximately $4,123 per pupil over four years.[9]

Overall, the results suggest that TIA not only raises student performance but does so at relatively low cost per unit of achievement gain, particularly in schools with high levels of vulnerability, where the program is most effective.

5.5 Are Designated Teachers Moving, and Where?

Previous sections have shown that the policy has reduced educational gaps between schools with different levels of vulnerability. However, the differences in monetary incentives across schools with different vulnerability levels not only create more incentive for teachers to stay in vulnerable schools, but also move to those schools, especially for those teachers who had already obtained a designation The policy's design offers larger monetary rewards for teaching in high-poverty and rural schools, theoretically creating an incentive for designated high-performing teachers to seek positions in more disadvantaged districts or schools.

I thus investigate whether designated teachers tend to move, and if so, whether they are moving to high-need districts (as opposed to leaving for more affluent districts or staying put). For this analysis, I track the subset of teachers who received a TIA designation and examine their mobility patterns in the years before and after designation, using non-designated teachers in the same districts as a comparison group in a teacher-level differences-in-differences framework represented by the following equation,

$$Y_{lt} = \sum_{k \neq -1} \beta_k \cdot \text{Designated}_{lt}^{(k)} + \mu_{D(l)t} + \zeta_l + \varepsilon_{lt} \tag{4}$$

where Y_{lt} is the outcome of teacher l in year t and $\text{Designated}_{lt}^{(k)}$ is a dummy variable taking the value of one if the teacher obtained a designation k years from t. The outcome is the vulnerability level of the school in which the teacher is in the concurrent year; its value increases if the teacher moves to a more vulnerable school. An increase in the vulnerability level would mean that the teacher moved to a more vulnerable school.[10] I include teachers' fixed effects ζ_l, and initial district-by-year fixed effects $\mu_{D(l)t}$, where I use the district in which the teacher was first in a district with TIA. This last district is also the level at which standard errors are clustered.

Event-time estimates from equation 4 are plotted in Figure 6. Figure 6a shows the difference in the average vulnerability of schools where designated and non-designated teachers work before and after designation. On average, after three years of designation, schools are scored as 0.5 points more vulnerable on a scale that goes from 0 to 4. In addition, Figure 6b shows that this effect is the strongest in people

[9] The authors report that *"a policy increasing spending by $1,000 per pupil for four years improves test scores by 0.031 s.d."* (Jackson and Mackevicius, 2024).

[10] The vulnerability score for one specific school might change from year to year. This means that a teacher might appear in a more vulnerable school if the school becomes more vulnerable too (without showing any teacher's migration). To avoid this plausible bias, I use the vulnerability level of the first implementation year, which is calculated for all schools.

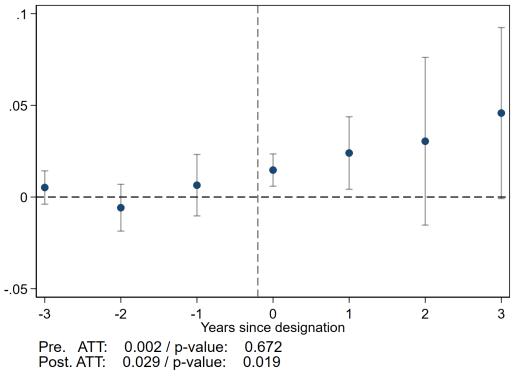
designated with the highest level ("master designation"). These teachers are also the ones who would benefit the most monetarily from changing to a disadvantaged school.

Do Departures of Designated Teachers Undermine School Performance?

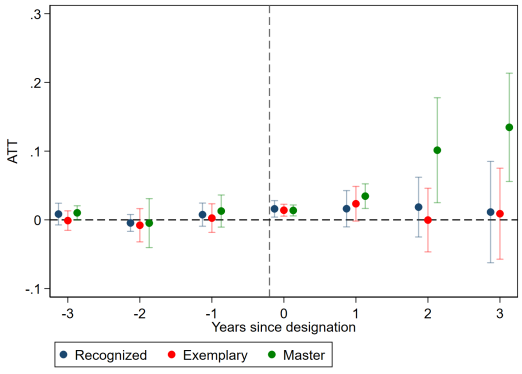
The finding that designated teachers tend to move to more vulnerable schools raises an important concern: if high-performing teachers are more likely to relocate to disadvantaged areas, this reassignment could come at the expense of the schools they leave behind. I examine whether the reallocation of designated teachers across schools generates unintended consequences for the campuses they leave behind. A detailed analysis can be found in Appendix B.9. These results show that schools that lose a designated teacher do not experience significant decline student achievement following the departure. If anything, effects are slightly negative in less vulnerable schools and weakly positive in more vulnerable ones, though none are statistically significant. Overall, the evidence suggests that TIA-induced teacher mobility does not produce measurable negative spillovers on student achievement in the schools that designated teachers leave behind.

Figure 6: Results on Movement of Designated Teachers

(a) Vulnerability Score of Designated Teachers



(b) Vulnerability Score by Designation



Note: Dynamic Treatment Effect calculated using conventional TWFE methodology. The outcome is the vulnerability score of the school where the teacher is teaching in year t, this score takes values from 0 to 4. Observations are at the teacher-by-school year level, and the sample consists of teachers in the districts under TIA when the district started the policy. It includes teachers' fixed effects and school district-by-year fixed effects. Standard errors are clustered at the level of school district from the first district in which the teacher is under TIA..

6 Conclusion

Texas's experience with the Teacher Incentive Allotment (TIA) demonstrates that a well-designed, statewide teacher performance pay program can yield positive impacts on achievement and teacher turnover. Our analysis finds meaningful improvements in student achievement: on average, standardized test scores increased by roughly 0.04 standard deviations under TIA. Importantly, these gains were largest in schools serving the most disadvantaged student populations, indicating that TIA not only raised overall performance but did it progressively, narrowing educational disparities across schools. Crucially, I find little evidence that these improvements came at the expense of higher-performing schools.

I also detect meaningful effects on the teacher workforce. Teacher turnover rates declined in TIA schools, translating into improved retention and slightly higher average teacher experience and tenure levels on those schools. Notably, I do not observe immediate changes in the composition of teachers' formal credentials, which implies that the short-run gains were driven more by keeping effective teachers in place rather than by any rapid influx of new qualifications (Papay and Kraft, 2015; Chetty et al., 2014a,b).

In addition, TIA induced a selective mobility response among high-performing teachers. After receiving a TIA designation (and the substantial bonus that comes with it), designated teachers became significantly more likely to transfer toward high-need schools, consistent with the program's incentive structure. These results are consistent with prior work showing that sizable and well-targeted policies can influence location choices (Papay et al., 2020).

In sum, TIA operated as both a retention policy and a targeted talent reallocation mechanism. It succeeded in keeping effective teachers on the job and in nudging those who do move after been designated to relocate to vulnerable schools. These behavioral responses are coherent with the program's design, resulting on achievement gains and reductions in the educational gap.

These results carry several clear policy implications. First, the design of the incentives appears crucial. The program's success seems to stem from rewarding excellence while explicitly channeling extra resources to disadvantaged settings. Offering the largest financial rewards for teaching in vulnerable schools is associated with stronger effects on those schools. If this positive relation comes from a causal effect on the incentive, then offering larger bonuses to talented teachers is an effective way of reducing educational gaps. If the largest effect in vulnerable schools is not due to the substantially higher bonus, but rather because there is more room for improvement in those places, then the funds allocated to those schools where incentives translate to larger student achievement.

Despite these encouraging findings, there are limitations to this early examination of the program. One limitation is that our outcome measures and time horizon are inherently limited to the program's initial years. I capture gains on standardized tests and near-term teacher movements, but we cannot yet observe longer-term outcomes for students -such as high school graduation, college enrollment, or later-life achievements-, or teachers, such as a sustained trajectory of teacher performance and effort beyond the short run. Future work should examine whether the test score improvements we document translate into lasting benefits for students, and whether designated teachers continue to excel over time.

Another consideration is the potential general equilibrium effect as the program matures. Our analysis exploits variation in adoption timing and compares early adopters to non-adopters; however, as TIA is scaled statewide (or if similar incentive schemes are adopted widely elsewhere), the competitive landscape for teachers will shift. In an equilibrium where the policy is scaled up to all school districts, some of the advantages seen in early-adopting districts might level out, and the movement of teachers between districts could reach a new balance. I expect this general equilibrium caveat to be limited, since the results suggest that most of the effect does not come from teachers' movements, but changes in productivity beyond movement patterns.

Thus far, I find no evidence that TIA caused adverse effects on non-targeted groups of students or on high-performing schools. However, performance pay programs can carry risks: for instance, teachers might focus excessively on tested subjects at the expense of broader learning, or morale could suffer among teachers who narrowly miss out on designations. Addressing questions about teachers' allocation of time, teachers' collaboration, or long-term student outcomes beyond test scores will be important as more data become available.

In conclusion, the first years of TIA offer a compelling case that aligning teacher incentives with student needs can be a powerful component of educational policies. By rewarding high performance and directing those rewards toward high-need schools, Texas improved outcomes in some of its most challenging educational settings while enhancing teacher retention and not disadvantaging higher-performing schools. If these trends persist, TIA will stand as an example of how decentralized, incentive-driven reform can boost both students' achievement and equality within a large public school system. For now, TIA's impacts suggest that thoughtful incentive policies can indeed motivate teachers in ways that benefit students, offering valuable lessons for other jurisdictions striving to raise achievement and reduce educational gaps.

References

Barrera-Osorio, F. and Raju, D. (2017). Teacher performance pay: Experimental evidence from pakistan. *Journal of Public Economics*, 148:75–91.

Bertrand, M., Burgess, R., Chawla, A., and Xu, G. (2020). The glittering prizes: Career incentives and bureaucrat performance. *The Review of Economic Studies*, 87(2):626–655.

Biasi, B. (2021). The labor market for teachers under different pay schemes. *American Economic Journal: Economic Policy*, 13(3):63–102.

Biasi, B., Fu, C., and Stromme, J. (2021). Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage. Technical report, National Bureau of Economic Research.

Burgess, S., Greaves, E., and Murphy, R. (2022). Deregulating teacher labor markets. *Economics of Education Review*, 88:102253.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230. Themed Issue: Treatment Effect 1.

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):2633–79.

Dal Bó, E., Finan, F., and Rossi, M. A. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly Journal of Economics*, 128(3):1169–1218.

Falch, T. (2011). Teacher mobility responses to wage changes: Evidence from a quasi-natural experiment. *American Economic Review*, 101(3):460–465.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from new york city public schools. *Journal of Labor Economics*, 31(2):373–407.

García, E. and Weiss, E. (2019). U.s. schools struggle to hire and retain teachers: The second report in "the perfect storm in the teacher labor market" series. Technical report, Economic Policy Institute.

Glewwe, P., Ilias, N., and Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 2(3):205–27.

Goodman, S. F. and Turner, L. J. (2013). The design of teacher incentive pay and educational outcomes: Evidence from the new york city bonus program. *Journal of Labor Economics*, 31(2):409–420.

Jackson, C. K. and Mackevicius, C. L. (2024). What impacts can we expect from school spending policy? evidence from evaluations in the united states. *American Economic Journal: Applied Economics*, 16(1):412–46.

Kirksey, J. J., Lansford, T., Crevar, A., and Mansell, K. (2024). From incentive to impact: The texas teacher incentive allotment's path to improved retention and achievement. working paper no. 02-001. *Center for Innovative Research in Change, Leadership, and Education*.

Leaver, C., Ozier, O., Serneels, P., and Zeitlin, A. (2021). Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from rwandan primary schools. *American Economic Review*, 111(7):2213–46.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., and Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from tanzania*. *The Quarterly Journal of Economics*, 134(3):1627–1673.

Muralidharan, K. and Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of Political Economy*, 119(1):39–77.

Nguyen, T. D., Pham, L. D., Crouch, M., and Springer, M. G. (2020). The correlates of teacher turnover: An updated and expanded meta-analysis of the literature. *Educational Research Review*, 31:100355.

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.

Papay, J. P., Taylor, E. S., Tyler, J. H., and Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, 12(1):359–88.

Podgursky, M. J. and Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, 26(4):909–949.

Rasul, I. and Rogger, D. (2018). Management of bureaucrats and public service delivery: Evidence from the nigerian civil service. *The Economic Journal*, 128(608):413–446.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J., McCaffrey, D. F., Pepper, M., and Stecher, B. M. (2011). Teacher pay for performance: Experimental evidence from the project on incentives in teaching (point). *Society for Research on Educational Effectiveness*.

7 Appendices

Appendix A - Designation details

The Teacher Incentive Allotment (TIA) is a Texas initiative that awards high-performing teachers one of three designation levels – Recognized, Exemplary, or Master – which correspond to significant annual stipend allotments for their district (ranging roughly from $3,000 up to $32,000 depending on designation and campus factors). Each TIA district must design a local designation system that evaluates teacher effectiveness based on two components: classroom teacher observation and student growth. (Districts may include additional optional measures like student surveys, teacher leadership roles, etc.) Each participating district's system must be approved by the Texas Education Agency (TEA), and designations are valid for five years. Teachers with National Board Certification automatically earn a Recognized designation.

Components

Teacher Observation: Most districts use the Texas Teacher Evaluation and Support System (T-TESS) rubric for the observation component. T-TESS includes four domains with 16 dimensions, scored 1–5 from Improvement Needed (1) to Distinguished (5). A teacher's observation score is typically the average of these eight dimensions' ratings. For example, Clyde CISD notes that the summative teacher observation score is the average across the eight T-TESS dimensions in Domains 2 & 3 (each dimension is worth up to 5 points). To be eligible for any designation, TEA requires a minimum rating of "Proficient" (level 3) on all 8 of these dimensions, this meas that no area can be below proficient.

Student Growth: Districts have flexibility in choosing valid student growth measures for each teacher group. Common growth measures include value-added models (VAM) using state STAAR exams, Student Learning Objectives (SLOs), pre- and post-tests, or portfolios. The student growth component for a teacher is usually expressed as the *"percentage of that teacher's students who met or exceeded an expected growth target"* on the chosen assessment. For example, TEA's statewide growth standard defines this as the percent of students who achieved or surpassed their predicted score (in a STAAR-based growth model). Districts ensure their growth measures are reliable and aligned to standards, and they train staff to administer and score them consistently.

Weighting of Observation vs. Growth in District Plans

Each TEA-approved local TIA plan must specify how the teacher's observation score and student growth results are **weighted** to produce an overall TIA composite score or designation score. The state allows districts to determine the weighting, provided both components are significant. In practice, districts use weightings ranging from about 20% up to 50% for the observation, with the remainder (50%–80%) for student growth. Here are some examples:

- Clyde CISD: Teacher Observation 20% / Student Growth 80%.
- Galena Park ISD: Teacher Observation 30% / Student Growth 70%.

- Lampasas ISD: Teacher Observation 30% / Student Growth 70%.
- Ingram ISD: Teacher Observation 40% / Student Growth 60%.
- Seguin ISD: Teacher Observation 50% / Student Growth 50%.

Despite varying weights, all plans must demonstrate a strong correlation between high observation ratings and high student growth – a validity check so that, for example, a teacher with stellar observations but low student growth (or vice versa) wouldn't be inappropriately designated. Some districts also incorporate a small percentage for "additional factors." For instance, Rice ISD's plan adds a fixed 10-point credit for high teacher attendance.

Eligible assignments

Each district can decide which assignments (subjects and grade levels) can be evaluated by the TIA system. Typically, districts choose core subjects for all grade levels. Here is an example of the teaching assignments and tools used for evaluation in Rice ISD:

Figure A1: Teaching Assignments and instruments, Rice ISD

Grade Level/Teaching Assignment	Teacher Observation	Who sets Expected Growth Target? How?	Student Growth (Pre-/Post-Tests)
Phase 1			
Pre-K	T-TESS	District (Graduated Percent Increase)	Circle/CLI
4th-8th Math and Reading, Algebra I, English I and English II (including SPED)	T-TESS	STAAR Transition Tables	STAAR
3rd Reading, 3rd Math, 5th Science, 8th Science, 8th SS, US History, Biology (including SPED)	T-TESS	District (Graduated Percent Increase)	BOY Released STAAR/EOY STAAR
K-2 Reading/Math (including SPED)	T-TESS	mCLASS	mCLASS
Intro to Culinary, Principles of Ag/Food/Natural Resources, Principles of Health Sci, Principles of Arts/Audio/Visual Tech	T-TESS	District (Graduated Percent Increase)	YouScience/ Precision Exams
SPED: Life Skills (Grades 4-11)	T-TESS	STAAR Transition Tables	STAAR-ALT
*All RCISD teachers eligible for designation through the National Board Certification			
Phase 2			
BIM	T-TESS	District (Graduated Percent Increase)	YouScience/ Precision Exams
Cosmetology, Physics, Spanish I, Chemistry, Algebra II, and Geometry	T-TESS	District (Graduated Percent Increase)	District-created

Source: Teacher Incentive Allotment (TIA), Teacher Overview 2025-26, Rice Independent School District.

Scoring Rubrics and Cut Score Thresholds for Designations

After combining the observation and growth components according to the chosen weights, teachers receive a composite TIA score (often on a 5-point scale or converted point system). Each district sets cut score thresholds to determine which designation (if any) a teacher earns. TEA provides statewide performance

standards as a reference, derived from statewide data, to guide these cut points. The statewide minimum standards for designation are:

- Recognized: ∼3.7 average on T-TESS (Domains 2&3) and 55% of students meeting growth targets.
- Exemplary: ∼3.9 average on T-TESS and 60% of students meeting growth.
- Master: ∼4.5 average on T-TESS and 70% of students meeting growth.

These figures correspond to roughly the top 33% (Recognized), 20% (Exemplary), and 5% (Master) of teachers statewide on each metric. Districts are not forced to use these exact numbers, but teachers must meet at least the minimum Proficient on all observation dimensions (as noted) and generally meet or exceed these growth percentages to be designated. In practice, many districts do adopt or stay close to the state's benchmarks. That said, some districts choose more stringent local criteria. For instance, Galena Park ISD fixed much higher student growth cut-offs: 70% of students growing for Recognized, 80% for Exemplary, and 90% for Master– far above the state minimums. This reflects a local decision to limit designations to truly exceptional growth outcomes in their context. Each district analyzes its own data distribution to finalize cut scores, and TEA allows that the percentage of teachers designated can vary by district. Score Calculation Examples: In TIA systems, a teacher's final score is calculated by applying the weights to their observation and growth results, then comparing to the cut scores. A simple example from Lampasas ISD: a teacher with a 4.48 T-TESS average and a 4.25 growth score (on a 5-point scale) would have a composite of $4.48 \times 30\% + 4.25 \times 70\% = 4.319$. This falls above the Exemplary cut (3.9) but below Master (4.5), resulting in an "Exemplary" designation recommendation.

Visualizing Performance Scoring and Designation Distribution

District TIA plans often include tables or scorecards to help teachers understand how the two components translate into a designation. These scorecards lay out the points or ranges needed. For instance, Rice CISD published a composite score chart: a Recognized teacher needed a total score ∼67–71, Exemplary ∼71–81, and Master∼ 81+ (out of 100). Those composite scores were the sum of Teacher Observation points (T-TESS average × 8), Student Growth points (% of students meeting growth × 50), plus any attendance points. In that model a teacher meeting the minimum Recognized criteria (3.7 avg and 55% growth) would score around 67 points total, just at the Recognized cut. These kinds of tables help visualize how each component contributes to the overall designation. Such breakdowns make it clear that a balance of strong teaching practices and strong student outcomes is required. If a teacher excels in observations but only a minority of students meet growth, they likely won't reach the designation cut. Conversely, high growth alone isn't sufficient without at least proficient-level instruction in the classroom (per the observation ratings requirement). By design, only teachers who demonstrate consistent excellence in both teaching practice and student progress earn TIA designations. Each approved district plan includes detailed rubrics and data validation steps to ensure the scoring is fair and credible.

Appendix B - Additional results

B.1 Determinants of designation status

Districts that adopt TIA and districts that do not are different. But *how different are designated teachers compared to non-designated teachers within TIA districts?* Answering this question provides us with a picture that will be useful to understand the effect that designation has on teacher turnover. In addition, it would help us to give validity to the designation process if more prepared teachers are more likely to receive a designation.

Table B1 reports the determinants of teacher designation across different levels of designation. Across specifications, advanced qualifications and accumulated teaching experience are strong predictors of designation status. Teachers holding a master's or doctoral degree are significantly more likely to be designated, and years of experience show a consistent and positive association with designation at all designation levels. These findings suggest that formal credentials and tenure in the profession remain central determinants of designation. This is reassuring and provides validity for districts' evaluation processes since these variables are typically associated with high-quality teachers.

The estimates also reveal important demographic patterns. Female teachers are significantly more likely than their male counterparts to receive designations across all designation levels. In contrast, black teachers are slightly, but systematically, less likely to be recognized than white teachers, while Hispanic teachers show a modest positive association with designation that is statistically significant at conventional levels in most models. Asian teachers, by contrast, do not show consistent differences.

Finally, Math and English teachers are consistently more likely to be designated than teachers of other subjects, whereas social science and science teachers show no difference. This subject-specific disparity is likely to reflect the fact that TIA districts start evaluating teachers in core subjects, and other subjects are included years later.

Overall, the results suggest that the program rewards observable teacher qualifications. However, it also generates uneven patterns of designation across gender, race, and subject areas.

Table B1: Determinant of each level of designation

	(1) Recognized designation	(2) Exemplary designation	(3) Master designation
Master/doc. degree	0.00195	0.00296	0.00117
	(0.00246)	(0.00212)	(0.00137)
Tenure	0.00344***	0.00247***	0.000828***
	(0.000604)	(0.000446)	(0.000185)
Experience	0.00112***	0.000630***	0.000269**
	(0.000228)	(0.000190)	(0.000115)
Asian	0.00558	0.000753	0.000401
	(0.00810)	(0.00501)	(0.00262)
Hispanic	0.00912**	0.00511*	-0.000619
	(0.00408)	(0.00277)	(0.00154)
Black	-0.00407	-0.00271	0.00108
	(0.00385)	(0.00330)	(0.00145)
Other ethnic	-0.00109	-0.00451	-0.00000924
	(0.00733)	(0.00598)	(0.00241)
Female	0.0121***	0.00814***	0.00137
	(0.00359)	(0.00277)	(0.00126)
Math teacher	0.0190***	0.0171***	0.00829***
	(0.00405)	(0.00366)	(0.00199)
Quality-proxi	0.00574*	0.00692**	0.00350**
	(0.00311)	(0.00275)	(0.00170)
Observations	51,920	51,920	51,920
School-level F.E.	Yes	Yes	Yes
Outcome Average	0.0594	0.0369	0.0112

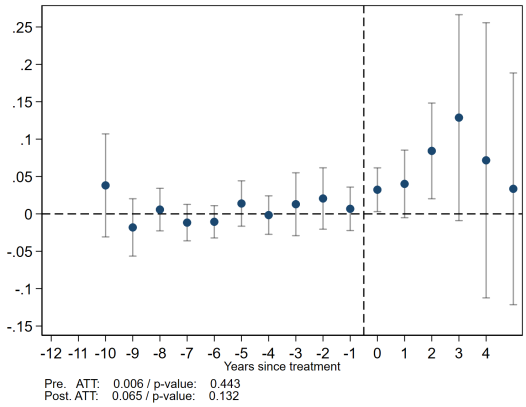
Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Columns 1 to 3 display the coefficients of an OLS regression of having at least a designation level in 2023 on potential determinants of designations. Outcomes in columns 1 to 3 reflect whether the teacher gets a recognized, exemplary, or master designation. Variables related to teacher appointment are approximations. I classify as an English teacher any teacher who registers English as her main teaching subject. Hence, different from the previous table, a teacher can be classified either as a math or english teacher in the same year at the same school, but not both.

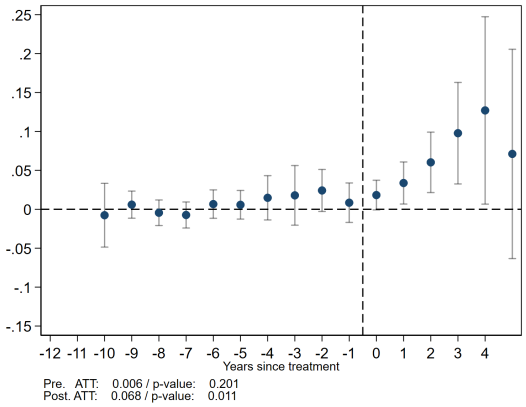
B.2 Results by subject

Figure B1: Results on test scores, math & reading separately

(a) Math



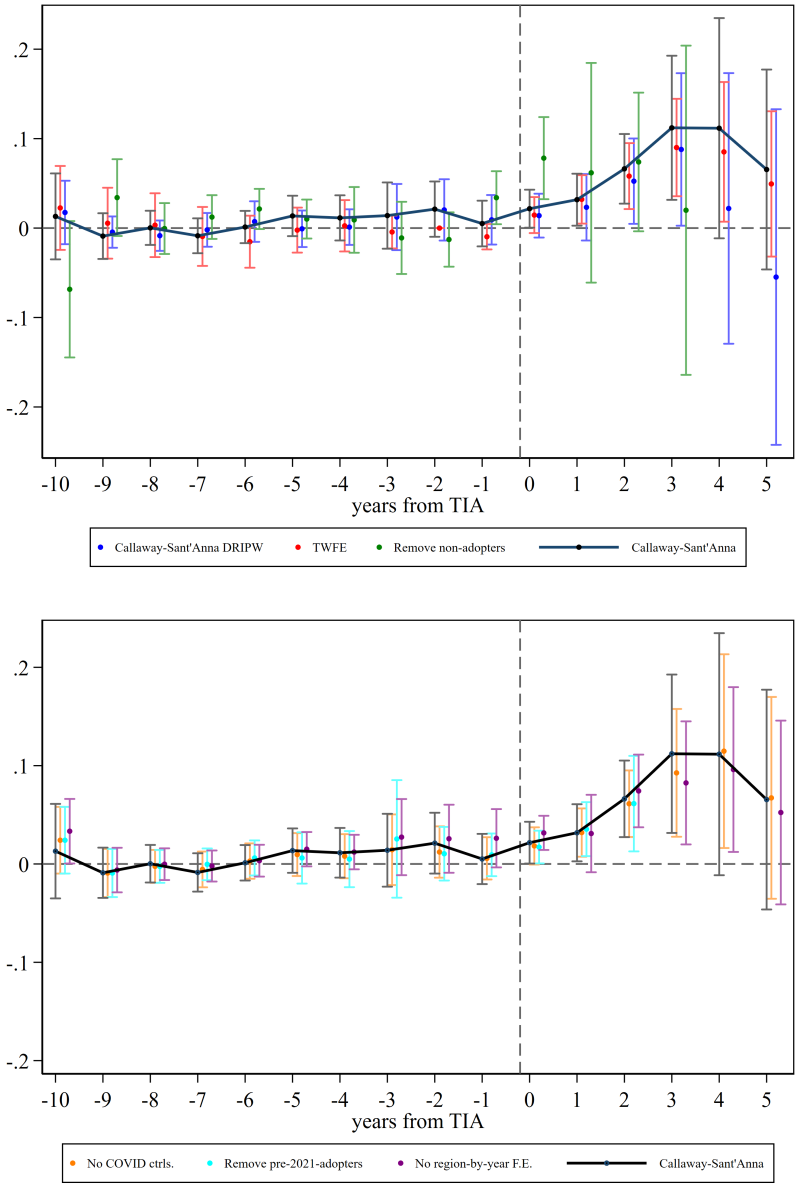
(b) Reading



Note: Dynamic effects calculated using Callaway and Sant'Anna (2021) estimation methods. Outcomes are standardized measures of math and reading scores of the STAAR test, a test implemented in all public schools between the 3rd and 8th grade levels between 2013 and 2024. It includes school-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

B.3 Results by method

Figure B2: Results on test scores, math & reading combined, several methods



Note: Dynamic effects calculated using several estimation methods. Outcomes are standardized measures of math and reading scores (pooled) of the STAAR test, a test implemented in all public schools between the 3rd and 8th grade levels between 2013 and 2024. It includes school-by-subject-by-grade level fixed effects, year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

B.4 Results by cohort of adoptions

Table B2: Group Average Treatment Effects on STAAR Test Scores

	(1) Total	(2) Math	(3) Reading
Group ATT	0.0408***	0.0468***	0.0346***
	(3.66)	(2.91)	(3.10)
G2019	0.0361	-0.00428	0.0755
	(0.61)	(-0.05)	(1.41)
G2021	0.0929**	0.131*	0.0553
	(2.14)	(1.90)	(1.47)
G2022	0.0504***	0.0606**	0.0401**
	(2.66)	(2.09)	(2.04)
G2023	0.0207	0.0172	0.0240
	(1.53)	(1.24)	(1.51)
G2024	0.0168	0.0260**	0.00752
	(1.40)	(2.17)	(0.50)
Observations	321,132	160,116	161,016
Nr. Untreated schools	2,250	2,250	2,248
Nr. Treated schools	2,469	2,467	2,468
Outcome Av. ctrl	0.029	0.014	0.044

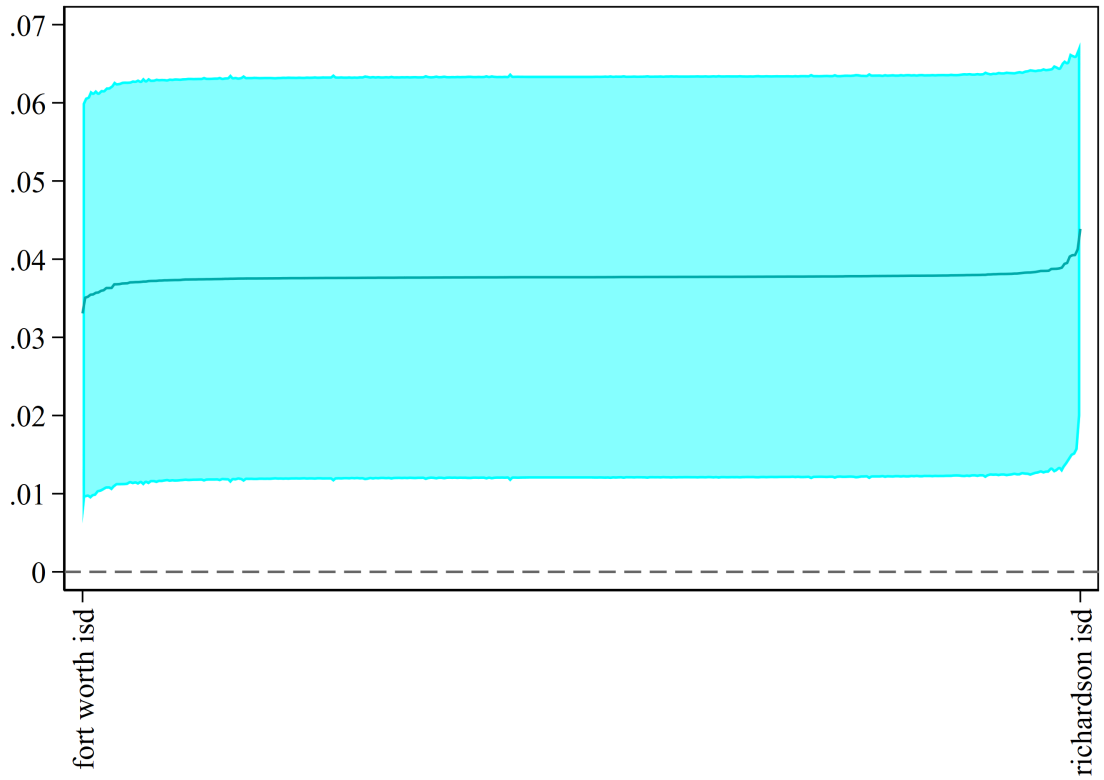
t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Group Average Treatment Effect calculated using Callaway and Sant'Anna (2021) estimation methods. Observations used in this regression are at the school year-by-school-by-subject-by-grade level, and I weight by the number of test takers. I use data between 2013 and 2024. It includes school-by-subject-by-grade level fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

B.5 Results Dropping one Treated District at a Time

Figure B3: Results Dropping one Treated District at a Time



This figure reports the estimated average treatment effects from 376 regressions, each excluding one treated school district in turn. This leave-one-out approach tests whether the results are driven by any single outlier district. All regressions follow the same specification as the main analysis. The estimated coefficients are plotted in ascending order, with the green line showing point estimates and the shaded cyan area indicating 95% confidence intervals. Standard errors are clustered at the school district level. Excluding the Fort Worth Independent School District yields the smallest effect, and excluding the Richardson Independent School District yields the largest coefficient.

B.6 Heterogeneity Analysis by Level of Experience

Figure B4: Exit rate by experience level

(a) Low experience (exp≤5) (b) Middle experience (5<exp≤10) (c) High experience (10<exp)

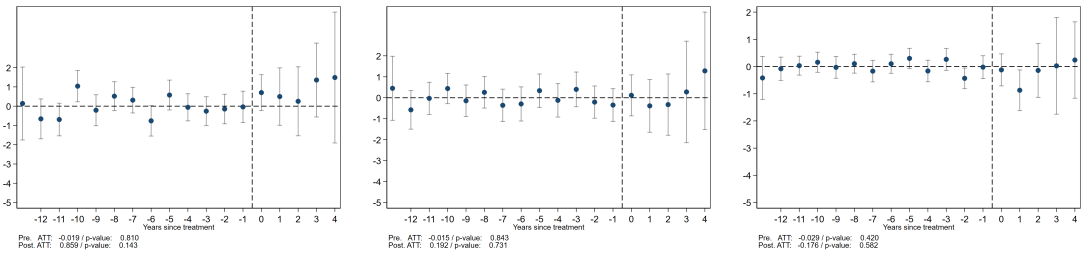
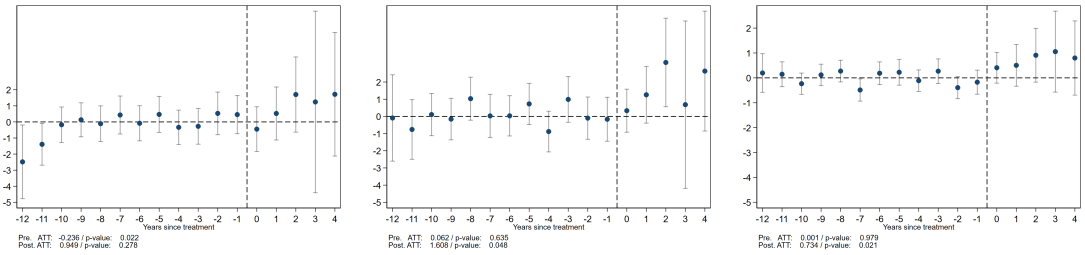


Figure B5: Entry rate by experience level

(a) Low experience (exp≤5) (b) Middle experience (5<exp≤10) (c) High experience (10<exp)



Note: Dynamic effects calculated using Callaway and Sant'Anna (2021) estimation methods. The outcomes measure the proportion of teachers within a given experience range who exit or enter a school, relative to the total number of teachers in that experience range. It includes school fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

Table B3: Group Average Treatment Effect by experience group

	(1) Low experience exit	(2) Middle experience exit	(3) High experience exit	(4) Low experience entry	(5) Middle experience entry	(6) High experience entry
Group ATT	0.615	0.0987	-0.133	0.523	0.900	0.570**
	(0.486)	(0.411)	(0.262)	(0.636)	(0.588)	(0.282)
Observations	80,318	80,318	80,318	80,318	80,318	80,318
Nr. Untreated schools	2,810	2,810	2,810	2,810	2,810	2,810
Nr. Treated schools	2,935	2,935	2,935	2,935	2,935	2,935
Outcome Av. ctrl	10	7	4	11	9	5

Standard errors in parentheses

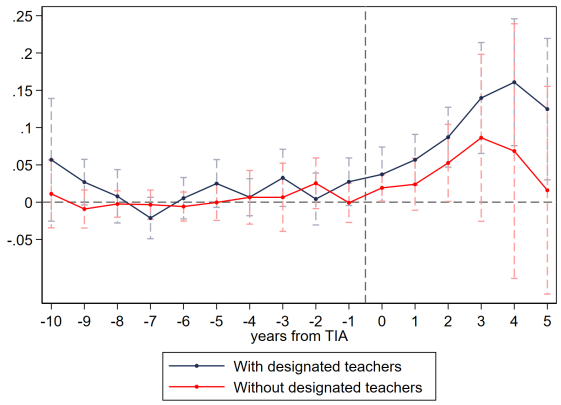
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Group Average Treatment Effects calculated using Callaway and Sant'Anna (2021) estimation methods. The outcomes measure the proportion of teachers within a given experience range who exit or enter a school, relative to the total number of teachers in that experience range between 2013 and 2023. It includes school fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

B.7 Results by existence of designated teacher or not

Figure B6: Results on test scores by designation in school or not

(a) Mathand Reading



Note: Dynamic effects calculated using Callaway and Sant'Anna (2021) estimation methods for two types of schools in adopting districts: schools that had at least one teacher who was designated, and schools that do not have any designated teacher. The outcomes measure the STAAR test scores normalized at the school-by-grade-by-subject-by-year level between 2013 and 2024. It includes school-by-grade-by-subject, region-by-year, rurality-by-year, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

Table B4: Group Average Treatment Effect by designation in school or not

	(1) With designated tch.	(2) Without designated tch.
Group ATT	0.0657***	0.0279***
	(0.0163)	(0.0106)
Observations	214,164	264,504
Nr. Untreated schools	2,250	2,250
Nr. Treated schools	846	1,623
Outcome Av. ctrl	0.145	0.145

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Group Average Treatment Effects calculated using Callaway and Sant'Anna (2021) estimation methods for two types of schools in adopting districts: schools that had at least one teacher who was designated, and schools that do not have any designated teacher. The outcomes measure the STAAR test scores normalized at the school-by-grade-by-subject-by-year level. It includes school-by-grade-by-subject fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Notice that, in contrast to conventional TWFE, Callaway and Sant'Anna (2021) calculates differences cohort by cohort, comparing scores with respect to not-yet treated observations. This makes it possible to calculate the difference between each pre-treatment period and the not-yet-treated, which means that no pre-treatment coefficient is deleted. Standard errors are clustered at the school district level. Standard errors are clustered at the school district level.

B.8 Methodology: Is the Impact Driven by Changes in Teacher Composition?

To what extent do these factors account for the observed improvements in test scores? In theory, we could disentangle this relationship using instrumental variables to isolate the exogenous variation in experience. In practice, however, finding instruments that shape teacher retention without simultaneously influencing student learning has proven particularly challenging. For instance, enhancing teachers' outside options could indirectly influence the local economy of students' parents, thereby violating the exclusion restriction. Moreover, other shift–share instruments used in turnover analyses often suffer from regression-to-the-mean concerns or limited statistical power. In this section, I explain two complementary approaches for assessing how changes in teacher composition affect test scores. The first constructs a measure of test performance that captures variation driven solely by changes in teacher composition, while the second uses a mediation analysis to quantify the extent to which composition explains the overall policy effect.

B.8.1 Details in the Effects on the Teacher-Explained Component of Achievement

To isolate the portion of outcome variation driven by teacher composition, I implement a two-step procedure. In the first step, I model test score variation left after accounting for the control variables and fixed effects in equation 1, using a comprehensive set of school-level covariates that capture the composition of teachers and their interaction with grade levels and subjects. These covariates are primarily indicator variables denoting whether school i in subject s employs teachers in specific demographic and professional categories. The categories include seven age groups, five tenure groups, five experience ranges, two

educational levels (postgraduate or not), and race categories, all interacted with grade-level dummies.

To ensure that composition measures are not themselves influenced by the policy, I estimate this model using only pre-treatment periods. I then apply a LASSO regression to select the most relevant predictors and reduce the risk of overfitting. This approach effectively identifies combinations of teacher characteristics most predictive of test score variation—for example, whether a school employs math teachers aged 35–40 with over five years of experience and two years of tenure. The predicted values from this model represent the component of student achievement attributable to variation in teacher composition, which I use to assess how much of the policy's total effect operates through compositional changes. Formally, the specification is as follows:

$$e_{istg} = \mathbf{subject_s} \times \mathbf{grade_g} \times \mathbf{X}'_{it}\Lambda + u_{istg} \qquad for\ t \leq 2018$$

where e_{istg} are the residuals of the regression 1 using years 2010 to 2018. In the second stage, I predict e_{istg} for all observations, both pre- and post-2019, whether they are adopters or not. I call this variable $\hat{Y}(teachers)_{istg}$, because it was meant to capture the variation in test scores that can be explained by characteristics of teachers in the school.

In the second step, I estimate the average treatment effect on the component of test scores explained by teacher composition as follows:

$$\hat{Y}(teachers)_{istg} = \theta \times Tretment_{D(i)t} + \epsilon_{istg}$$

where $Tretment_{D(i)t}$ represents the TIA treatment variable that captures the variation identifying the effect on test scores in equation 1. Specifically, it reflects the residual variation in the district-level implementation of TIA after accounting for the controls and fixed effects included in equation 1. This regression isolates how much of the policy's estimated impact on student achievement can be attributed to observable changes in teacher composition, as captured by $\hat{Y}(teachers)_{istg}$.

Table 6 presents the results. The outcomes are, in column 1, standardized STAAR test scores and, in column 2, $\hat{Y}(teachers)_{istg}$, the student achievement can be attributed to observable changes in teacher composition. Notice that the effect shown in Table 7, column 1, is smaller than that in our primary analysis. This is because I excluded 2024 from the study, as I only have teachers' data up to 2023. The second column reports the estimated effect on the compositional component of test scores, alongside the average treatment effect on overall test scores shown in the first column. Because teacher-level variables are available only through 2023, the estimated effect on overall test scores in column one is slightly smaller than in previous specifications that included 2024 data. The results indicate that changes in teacher composition at the school level do not account for a statistically significant share of the policy's impact on test scores. If anything, the magnitude is modest—explaining roughly 6% of the total effect (0.00187 / 0.0316).

B.8.2 Details on Mediation Analysis

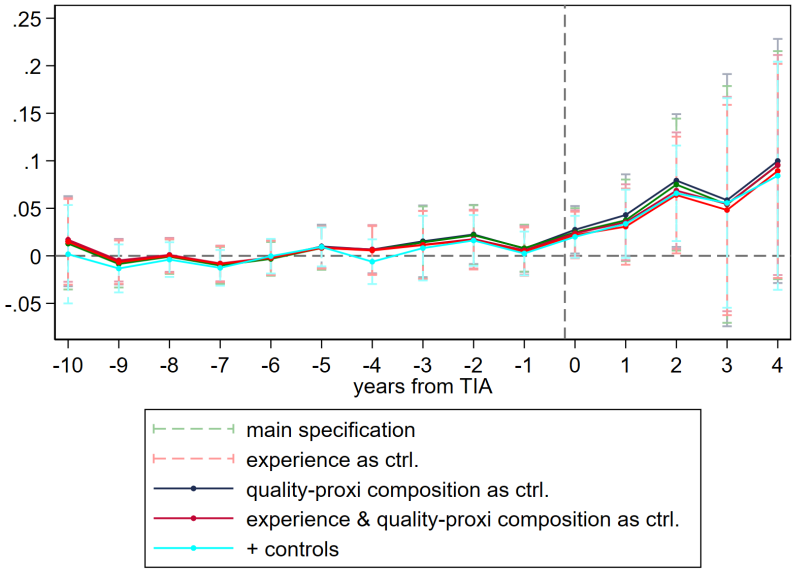
To better understand the mechanisms driving the observed improvements in student achievement, I conduct a mediation analysis that decomposes the total effect of TIA into direct and indirect components. Specifically, I include observable mediators related to teacher composition—such as average experience, retention, and turnover—that may channel part of the policy's impact on test scores. The coefficient on the policy dummy then captures the portion of the total effect that remains after accounting for these mediators. This channel likely reflects other mechanisms- such as increased teacher effort, improvements in school administration, or enhanced management practices- that are not directly observed in the data. It is important to note that the validity of this analysis relies on strong identification assumptions, such as the absence of omitted variables that jointly affect both the mediators and the outcome. Given these requirements, the results should be interpreted as suggestive evidence on the relative importance of different channels.

Table 7 presents the total effect that remains after accounting for these mediators. Notice that the effect shown in Table 7, column 1, is smaller than that in our primary analysis because this analysis excludes 2024, as teacher-level data are only available through 2023. Approximately 2% of the achievement gain – calculated as the percentage change between the coefficient of column 1 and 2, i.e. $(1 - \frac{0.0309}{0.0316})$ – is explained by changes in teacher experience. Column (3) introduces changes in $\hat{Y}(teachers)$, the observed change in teachers' characteristics represented in test scores values estimated with observations prior to 2019, as a mediator. Changes in scores estimated by teachers' composition account for roughly 3% of the total effect. Column (4) includes both mediators. Together, they explain about 6% of the total effect, a share close to the sum of the individual contributions. The final columns reflect a more flexible specification. Here, I add controls for teacher flows (entries and exits) and teachers' demographic composition (e.g., number of teachers by race), and interact these mediators with subject and grade dummies. Interestingly, the mediated share of the total effect increases to 14%.

Two lessons follow from this analysis. First, teacher experience and composition do not seem to explain a substantial part of test scores, despite controlling for characteristics in a very flexible way. Second, the majority of the effect is attributable to factors not captured by observable workforce shifts, likely direct responses to incentives, such as greater instructional effort or administrative adaptations.

In summary, I find no evidence that observable changes in the teacher workforce explain a substantial share of the overall effect. At best, the most generous estimates suggest that shifts in teachers' characteristics account for no more than 14% of the total improvement in test scores. This points to other channels—beyond observable composition—through which the policy may have increased student performance, such as greater teacher effort or improvements in school management practices. Interestingly, this 14% of the effect explained by compositional changes of the teacher workforce is close to the 1/5 portion explained by selection by Leaver et al. (2021).

Figure B7: Mediation results on test scores, math & reading combined



Note: Dynamic effects calculated using Callaway and Sant'Anna (2021) estimation methods. Each line represents a specification that includes additional mediators as control variables. The outcomes measure the STAAR test scores normalized at the school-by-grade-by-subject-by-year level between 2013 and 2023. It includes school=by-grade-by-subject fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level.

B.9 Results by existence of designated teacher or not

The finding that designated teachers tend to move to more vulnerable schools raises an important concern: if high-performing teachers are more likely to relocate to disadvantaged areas, this reassignment could come at the expense of the schools they leave behind. In this section, I examine the consequences of losing a designated teacher.

To do so, I track schools participating in the TIA program and identify whether they employed a designated teacher and whether that teacher subsequently left the school. I restrict the sample to schools that have had at most one designated teacher, ensuring that when the designated teacher departs, all remaining students are now taught exclusively by non-designated teachers.

I then estimate the change in test scores before and after the departure of a designated teacher, comparing schools that lost such a teacher to those that did not. This difference-in-differences design allows me to identify the effect of losing a designated teacher on student achievement. Formally, I estimate the following specification:

$$Y_{isgt} = \beta_0 + \beta_1 TeacherLeft_{it} + \beta_2 \mathbf{X}_{it} + \delta_{isg} + \eta_t + \epsilon_{isgt} \qquad t \geq startyear$$

where Y_{isgt} is the test score at school i, subject s, grade level g in year t. $TeacherLeft_{it}$ is a dummy variable that takes the value of one if the school lost a designated teacher any year before t. I use only years where the school is under the TIA program ($t \geq startyear$). Importantly, this effect considers both

the direct effect of the teacher in its students and the potential spillover of leaving the school in other students. Because of this reason, the effect is potentially biased towards zero.

Table B5 shows that the departure of designated teachers does not substantially harm student achievement after they leave. If anything, the effects are slightly negative in less vulnerable schools and weakly positive in more vulnerable ones. This pattern could arise if designated teachers generate spillovers before leaving or if their departure signals advancement opportunities for other teachers. However, none of these effects are statistically significant at conventional levels.

Moreover, the potential impact of losing a designated teacher should depend on school size. In larger schools, designated teachers are less likely to teach a substantial share of students, which reduces the expected effect of their departure and accentuate attenuation bias. Conversely, if losing a designated teacher truly harmed performance, the effect should be stronger in smaller schools. Contrary to this expectation, Columns (3) and (4) of Table B5 show that the estimated effect becomes slightly more negative in larger schools, reinforcing the conclusion that the loss of a designated teacher does not meaningfully reduce test scores.

Table B5: Average Effect of Designated Teacher Leaving School on STAAR Math and Reading tests combined

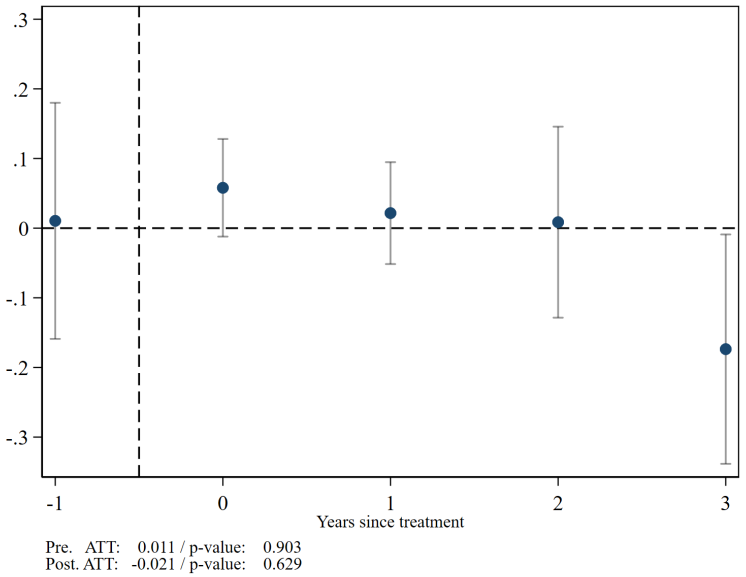
	(1)	(2)	(3)	(4)
Lost Designated	0.0260	-0.0263	0.186***	0.0727
	(0.0251)	(0.0313)	(0.0516)	(0.0946)
Lost Designated × Vulnerability score		0.0343*		0.0447*
		(0.0198)		(0.0241)
Lost Designated × Size quintile 2			-0.195***	-0.167***
			(0.0602)	(0.0571)
Lost Designated × Size quintile 3			-0.153**	-0.117
			(0.0606)	(0.0771)
Lost Designated × Size quintile 4			-0.147**	-0.0760
			(0.0579)	(0.0764)
Lost Designated × Size quintile 5			-0.167***	-0.129*
			(0.0563)	(0.0708)
Ctrls. × year	✓	✓	✓	✓
School × subject × grade level F.E.	✓	✓	✓	✓
District × year F.E.	✓	✓	✓	✓
Observations	23,874	23,874	23,874	23,874
Nr. Untreated schools	1,443	1,443	1,443	1,443
Nr. Treated schools	26	26	26	26
Outcome Av. ctrl	-0.128	-0.128	-0.128	-0.128

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Averate Treatment Effect of losing a designated teacher on test scores. The outcome is the STAAR test score for math and reading, normalized, for each grade level between 3rd and 8th, for each public school. The sample includes only schools in TIA districts since the year each district first implemented TIA. "Lost Designated" is a variable indicating whether the school lost a designated teacher before year t. Regressions include school-by-grade-by-subject fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the school district level. The second column evaluates whether losing a designated teacher might affect vulnerable schools; the results show that, if so, the opposite is true. Column three examines interactions between losing a designated teacher and the size of the school, measured by the number of teachers. The last column includes both interactions.

Figure B8: Dynamic Effect of Designated Teacher Leaving School on STAAR Math and Reading tests combined

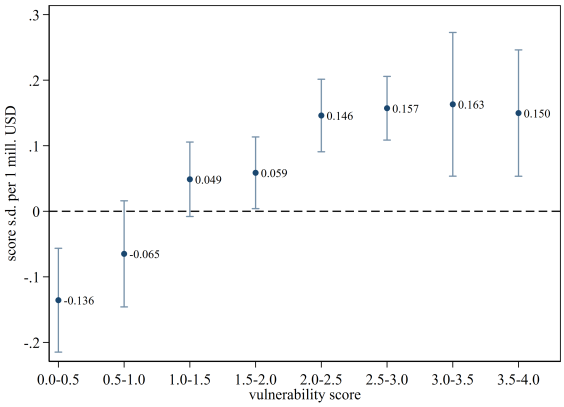


Dynamic effect of losing a designated teacher on test scores. The outcome is the STAAR test score for math and reading, normalized, for each grade level between 3rd and 8th, for each public school. The sample includes only schools in TIA districts since the year each district first implemented TIA. The x-axis shows the time passed since the school lost a designated teacher; 0 indicates it happened that year. Regressions include school-by-grade-by-subject fixed effects, region-by-year fixed effects, rurality-by-year fixed effects, and COVID policies-by-year fixed effects. Standard errors are clustered at the

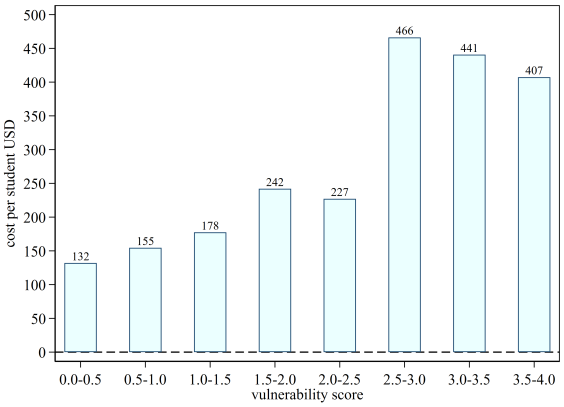
B.10 Cost Effectiveness

Figure B9: Effects and Costs by levels of vulnerability over four years into the program

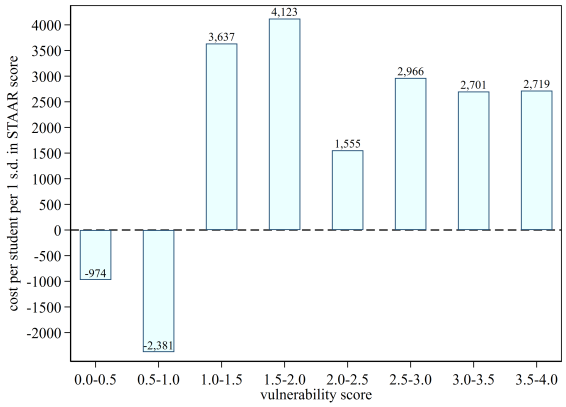
(a) Effect on STAAR in s.d.



(b) Cost per Student



(c) Cost per Student per 1 s.d.



The first graph shows the heterogeneous average treatment effect of TIA over the first three years for eight bins of vulnerability score, using schools in districts that have been in the program for 4 years or more. The outcome is at the school-by-by-subject-by-year-by-grade level. I use school-by-subject-by-grade-level fixed effects, region-by-year fixed effects, and Covid variables interacted with year. Standard errors are clustered at the school district level. The second graph shows, for each of the 8 bins of vulnerability, the allotment paid by the states to schools at each level of vulnerability during the first 4 years, divided by the number of students enrolled, for schools in districts that have been in the program for 4 years or more. The third graph shows the division between the values in the second and first graphs, providing the final estimate of the average cost per 1 s.d. increase in test scores per student over three years in the program.