education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 34 Number 80 September 15, 2026 ISSN 1068-2341

Can Teacher Pay-for-Performance Programs be Successful? The Case of the Texas Teacher Incentive Allotment

Jaehoon Lee
Michael Strong
Kristin Mansell
&
Heather Greenhalgh-Spencer
Texas Tech University
United States

Citation: Lee, J., Strong, M., Mansell, K. & Greenhalgh-Spencer, H. (2026). Can teacher pay-for-performance programs be successful? The case of the Texas Teacher Incentive Allotment. *Education Policy Analysis Archives, 34*(80). https://doi.org/10.14507/epaa.34.9573

Abstract: Pay-for-performance (PFP) programs in public education are controversial, and evaluations tend to focus on outcomes even though success also depends on design and implementation. We examine Texas' Teacher Incentive Allotment (TIA), a large state-funded PFP program, using a three-part framework—design validity, implementation fidelity, and outcome utility. Drawing on administrative data from 79,198 teachers across 338 districts, we ask whether designations align with effectiveness, whether districts apply metrics consistently, and whether designation is associated with performance. Combining growth and observation measures, only 3% of teachers were designated despite below-threshold scores, and 8% not designated despite above-threshold scores. Also, designation aligned with performance (AUC = .95), indicating that districts largely applied state guidance appropriately. Designation was also associated with improved or sustained performance, though our observational design cannot establish causation. TIA illustrates how a large-scale PFP program can combine state rigor with local flexibility, with implications for policymakers weighing merit-pay reforms.

Journal website: https://epaa.asu.edu/ojs/
Facebook: /EPAAA
Bluesky: @epaa.bsky.social

Manuscript received: 8/8/2025
Revisions received: 23/6/2026
Accepted: 10/7/2026

Keywords: performance pay; teacher incentive allotment; program evaluation; validity; value-added models; Texas

¿Pueden tener éxito los programas de remuneración docente basada en el desempeño? El caso del *Texas Teacher Incentive Allotment*

Resumen: Los programas de remuneración basada en el desempeño (PBD) en la educación pública son objeto de controversia, y sus evaluaciones suelen centrarse en los resultados, aunque su éxito también depende del diseño y la implementación. En este estudio examinamos el *Texas Teacher Incentive Allotment* (TIA), un programa estatal de gran escala financiado con fondos públicos para remunerar el desempeño docente, utilizando un marco analítico de tres componentes: validez del diseño, fidelidad de la implementación y utilidad de los resultados. A partir de datos administrativos de 79.198 docentes pertenecientes a 338 distritos escolares, analizamos si las designaciones se corresponden con la efectividad docente, si los distritos aplican de manera consistente las métricas establecidas y si la designación se asocia con el desempeño. Al combinar medidas de progreso académico y observaciones de aula, encontramos que solo el 3 % del profesorado recibió una designación pese a obtener puntuaciones inferiores al umbral establecido, mientras que el 8 % no fue designado a pesar de superar dicho umbral. Asimismo, la designación mostró una fuerte correspondencia con el desempeño (AUC = .95), lo que indica que los distritos aplicaron, en términos generales, las directrices estatales de manera adecuada. La designación también se asoció con un mejor desempeño o con el mantenimiento de niveles elevados de desempeño, aunque el diseño observacional del estudio no permite establecer relaciones causales. El TIA ilustra cómo un programa de remuneración basada en el desempeño implementado a gran escala puede combinar el rigor estatal con la flexibilidad local, ofreciendo implicaciones relevantes para responsables de políticas que evalúan reformas de remuneración por mérito.

Palabras-clave: remuneración basada en el desempeño; *Teacher Incentive Allotment*; evaluación de programas; validez; modelos de valor agregado; Texas

Programas de remuneração docente por desempenho podem ser bem-sucedidos? O caso do *Texas Teacher Incentive Allotment*

Resumo: Os programas de remuneração por desempenho (RPD) na educação pública são controversos, e suas avaliações tendem a se concentrar nos resultados, embora seu sucesso também dependa do desenho e da implementação. Neste estudo, examinamos o *Texas Teacher Incentive Allotment* (TIA), um programa estadual de grande escala financiado com recursos públicos para remuneração por desempenho docente, utilizando um referencial analítico composto por três dimensões: validade do desenho, fidelidade da implementação e utilidade dos resultados. Com base em dados administrativos de 79.198 professores distribuídos em 338 distritos escolares, investigamos se as designações correspondem à eficácia docente, se os distritos aplicam as métricas de forma consistente e se a designação está associada ao desempenho. Combinando medidas de crescimento da aprendizagem e observações de sala de aula, constatamos que apenas 3% dos professores receberam designação apesar de apresentarem pontuações abaixo do limite estabelecido, enquanto 8% não receberam designação apesar de ultrapassarem esse limite. Além disso, a designação apresentou forte associação com o desempenho (AUC = .95), indicando que os distritos, em sua maioria, aplicaram adequadamente as diretrizes estaduais. A designação também esteve associada à melhoria ou à manutenção do desempenho docente, embora o delineamento observacional do estudo não permita estabelecer relações de causalidade. O TIA demonstra como um programa de remuneração por desempenho

em larga escala pode combinar rigor estadual com flexibilidade local, oferecendo implicações relevantes para formuladores de políticas que avaliam reformas de remuneração por mérito.
Palavras-chave: remuneração por desempenho; *Teacher Incentive Allotment*; avaliação de programas; validade; modelos de valor agregado; Texas

Can Teacher Pay-for-Performance Programs be Successful? The Case of the Texas Teacher Incentive Allotment

Over the past 30 years, education policymakers across the United States and globally have grappled with the challenge of improving teacher effectiveness, often through Pay-for-Performance (PFP) programs. In U.S. school districts, PFP programs increased from a 7.9% share in 2004 to 11.3% in 2012 (Imberman, 2015), largely catalyzed by federal funding through the Teacher Incentive Fund (TIF), which ran from 2006 to 2012. The policy rationale behind PFP, grounded in expectancy theory (Vroom, 1964), is clear: reward high-performing teachers to enhance student achievement and attract talent to high-need schools. Although this rationale runs through teacher motivation, our study measures observed performance rather than motivation itself; we use the term in its expectancy-theory sense and treat it as one interpretive lens, distinct from the intrinsic and extrinsic framing of self-determination theory (Ryan & Deci, 2000) that we develop below.

Researchers suggest that this interest in PFP programs has continued since TIF (e.g., Springer & Taylor, 2016). Previously, more often referred to as merit pay (and we use the terms PFP and merit pay interchangeably in this article), these programs variously attempted to reward teachers with additional compensation based on their effectiveness at raising student achievement or by other ways demonstrating superior teaching skills. In this way, they represent a policy response aimed at improving accountability by strengthening the relationship between teacher rewards and student learning (Kirksey et al., 2024). Proponents of PFP programs argue that they incentivize high performance by offering teachers increased pay and publicly acknowledging their contributions to student success. This incentive, of more money and greater acknowledgment, is thought to improve student achievement (Milanowski, 2003). They further contend that PFP programs attract and retain high-quality teachers and encourage them to engage in professional development, thereby qualifying them for effectiveness-based financial incentives (Loeb & Myung, 2020). Arguments against PFP programs suggest that merit pay encourages teaching to the test, fails to consider contextual factors such as student socio-economic status, leads to biased evaluations, has a negative effect on collaboration, may misclassify teachers for rewards, and may undermine rather than inspire motivation (Jacob, 2005; Jacob & Levitt, 2003). Both major unions, the American Federation of Teachers (AFT) and the National Education Association (NEA) oppose such programs, objecting in particular to compensation tied to student test scores—which they view as invalid and unreliable proxies for teacher quality—and warning that performance pay can erode collegial collaboration, invite subjective or inequitable evaluation (Murnane & Cohen, 1986), and prove fiscally unsustainable once external funding lapses (West & Mykerezi, 2011).

Despite conceptual appeal, many PFP initiatives have faltered due to design flaws, implementation inconsistencies, and limited evidence of sustained impact. Against this backdrop, Texas' Teacher Incentive Allotment (TIA) represents a critical test case for policymakers seeking scalable, durable strategic compensation reform. This study contributes to policy and research by analyzing the TIA's design, implementation, and early effects on teacher motivation, as defined by Expectancy Theory, and offers one of the most comprehensive, data-driven evaluations of a state-led PFP initiative to date.

The Logical Basis for our Study

The policy logic of PFP rests on a motivational premise that we think is worth making explicit. Vroom's (1964) expectancy theory holds that the motivational force behind an individual's effort is a function of three beliefs: *expectancy*, the perceived likelihood that increased effort will produce higher performance; *instrumentality*, the perceived likelihood that higher performance will in turn yield a valued outcome; and *valence*, the subjective value the individual attaches to that outcome. Because the three combine multiplicatively, motivation collapses if any one approaches zero: a teacher who doubts that effort can raise measured performance, who distrusts that strong performance will actually be rewarded, or who places little value on the reward will not be moved to greater effort, regardless of the other two.

We use the term motivation in the sense intended by expectancy theory—the force directing a teacher's effort toward a valued outcome—rather than in the sense developed in self-determination theory, where motivation is differentiated into intrinsic and extrinsic forms with distinct antecedents (Ryan & Deci, 2000). Our data do not measure motivation directly in either sense; we observe teachers' performance over time and treat motivation as one interpretive lens on those patterns. Whether external rewards such as performance pay enhance or crowd out intrinsic motivation, a central question in the self-determination literature, lies beyond what our data can address.

This framing clarifies what a PFP program must accomplish to influence behavior, and it anticipates the framework we develop below. Instrumentality, which is the cornerstone of any incentive system, depends directly on whether designation decisions are valid and consistently applied: if the metrics used to award designations do not track teacher effectiveness (a design problem) or are applied erratically across districts (an implementation problem), teachers cannot reasonably believe that performing well will yield the reward, and the incentive is lost. Valence depends on whether teachers regard the bonus and its associated professional status as worth pursuing, the concern at the heart of the program buy-in literature (Imberman & Lovenheim, 2015). Expectancy depends on the belief that teachers can raise their own measured performance, a belief shaped in turn by how transparent and attainable the designation criteria appear. Expectancy theory thus binds our design, implementation, and outcome questions into a single motivational assertion: a PFP program changes behavior only when it is designed and implemented so that effort, measured performance, and valued reward are reliably connected in teachers' experience.

Our framework follows the logic of program theory and the standard sequence of program evaluation, which treats the quality of a program's *design*, the fidelity of its *implementation*, and the value of its *outcomes* as separable objects of evaluation (Rossi et al., 2019). A program may be well designed yet poorly implemented or faithfully implemented yet built on a flawed design; each stage is a necessary condition for the next, so failure at any stage constrains what later stages can achieve. We therefore treat design, implementation, and outcomes as three sequential levels at which a PFP program demonstrates success or falls short. Figure 1 represents our own synthesis of this evaluation logic applied to the design, implementation, and outcome elements of PFP programs identified in the literature.

Literature Review

Program Design

Designing a successful PFP program, beyond funding attractive bonuses, begins with how teachers are screened and selected for merit pay. Programs funded under the 2009 American

Recovery and Reinvestment Act generally required that effectiveness be judged on both student-achievement growth and classroom observation, on the theory that incentives would improve the workforce both by sorting (i.e., attracting and retaining higher performers while encouraging weaker ones to leave (the composition effect; Glazerman et al., 2011) and by motivating individual teachers to improve (the incentive effect).

The literature offers no consensus on optimal design but identifies recurring choices. Incentives may be individual- or group-based: group rewards risk free-riding (as in Houston's now-defunct ASPIRE program; Imberman & Lovenheim, 2015), while individual rewards can erode collaboration. Performance is typically captured with a student-growth metric (preferred over raw scores to avoid favoring teachers of high-ability students), usually paired with an observational evaluation. The incentive structure then takes one of three forms: target-level systems rewarding a fixed benchmark (e.g., Nashville; Springer et al., 2010), rank-order tournaments rewarding relative standing, and piece-rate systems rewarding each increment of gain. Because only tournaments operate within a fixed budget, administrators tend to favor them; some programs blend approaches, as in the "pay-for-percentile" design (Barlevy & Neal, 2012; Fryer et al., 2012).

Teacher support for PFP is mixed and appears to depend on both program design and teacher characteristics. Several studies have found teachers generally favorable toward PFP in principle (Azordegan et al., 2005; Ballou & Podgursky, 1993; Heneman & Milanowski, 1999; Milanowski, 2007), and evaluations of specific programs—including Denver's ProComp, the Texas Educator Excellence Grant, and the TIA itself—report broadly consistent support (Briggs et al., 2014; Lee et al., 2021; Springer et al., 2011). That support is conditional, however: teachers are more positive when they help design the system and when rewards are tied to evaluation rather than to test scores alone (Cornett & Gaines, 1994; Farkas et al., 2003), with approval falling as low as 17% in some settings (Goldhaber et al., 2011). Reported support also varies systematically with teacher characteristics, tending to be higher among novice than veteran teachers, among secondary than elementary teachers, and among male than female teachers (Ballou & Podgursky, 1993; Farkas et al., 2003; Goldhaber et al., 2011; Hill & Jones, 2020; Jacob & Springer, 2007). Taken together, this work suggests that buy-in is less a fixed attribute of teachers than a product of how a program is designed and communicated, an inference consistent with expectancy theory's emphasis on valence. This article adds to and extends this literature by sharing data around teacher buy-in for the Texas TIA program.

Program Implementation

Little research has focused specifically on program implementation, particularly on the use of metrics for designating teachers. Some studies have included information on teachers' retrospective perceptions of fairness. Cornett and Gaines (1994) suggested that teachers' perceptions of unfairness were largely responsible for the failure of PFP programs across the country. Lee et al. (2021) found that early participants in the TIA program generally perceived it to be fair (especially teachers who received designations!) but high school teachers were often less supportive and perceived it as less fair than teachers at other grade levels. It is not clear why this discrepancy should exist.

Teachers and administrators may differ in their perceptions of PFP programs (e.g., Reddy et al., 2018). Teachers may feel demoralized or negatively pressured by poorly implemented systems with components that are difficult to understand (e.g., Bradford & Braaten, 2018), or by systems that prioritize accountability over teacher development (e.g., Deneire et al., 2014; Ford & Hewitt, 2020). In short, PFP programs are more likely to be successful if they are implemented in a way that produces teachers' perceptions of fairness because they have a full understanding of the program's components and its potential impact (King & Paufler, 2020; Reddy et al., 2018).

The Dallas Independent School District's Teacher Excellence Initiative (TEI) (an immediate precursor to the TIA that likewise tied compensation to a combination of student growth, classroom observation, and stakeholder feedback) has generated the most direct, though conflicting, evidence on this family of programs. Using synthetic control methods to compare Dallas with similar Texas districts, Hanushek et al. (2023) attributed gains of roughly 0.2 standard deviations in mathematics and 0.1 in reading to the district's evaluation and pay reforms. Examining the same program at the student level from 2012 to 2019, however, Wang et al. (2025) found the association between TEI and achievement inconclusive, with effects varying across student subgroups. Those two careful analyses of one program reach divergent conclusions, underscoring how sensitive estimated PFP effects are to method and unit of analysis, a caution we carry into the interpretation of our own results.

Program Outcomes

Research on the effects of PFP on student learning is extensive but inconclusive. A meta-analysis of 37 studies (26 in the United States) found mixed results leaning toward a small positive effect (Pham et al., 2021). World Bank–funded reviews were less encouraging: most of the 15 evaluations from low- and middle-income countries showed no learning gains, and only four of the 28 high-income-country programs did (Breeding et al., 2021). Cross-national PISA comparisons, by contrast, suggested roughly a quarter-standard-deviation advantage in math, science, and reading for countries with PFP (Woessmann, 2011), while Imberman's (2015) review of 11 studies again found mixed results.

Across these reviews, design variation is repeatedly cited as a source of divergent findings. Imberman (2015) argued for incentives based on multiple outcomes rather than test scores alone, and against the rank-order tournaments common in U.S. programs, favoring piece-rate designs. Two cases illustrate the stakes: the Chicago Heights pay-for-percentile program, which paid teachers in advance under contracts requiring repayment if students did not improve, produced gains approaching a standard deviation (Fryer et al., 2012)—an effect driven by loss aversion that is hard to recommend for general use—whereas Nashville's high 85th-percentile threshold yielded no math gains (Springer et al., 2010).

Effects on recruitment and retention are studied less often; of 14 such studies in Pham et al. (2021), six showed improvement in teacher retention and recruitment outcomes, five were inconsistent, and three were null. Whether gains persist after incentives end, and whether they attract genuinely higher-quality teachers, remains unclear—one program improved recruitment chiefly by attracting more money-oriented candidates (Breeding et al., 2021). Findings on motivation are comparatively encouraging: Pham et al. (2021) reported that programs with positive test-score effects generally showed motivation gains as well, with stronger effects internationally and a lone exception in a small Arkansas pilot, though one U.S. study found PFP did not change teachers' motivation, instruction, or collegiality (Yuan et al., 2013).

Based on the literature reviewed here, we conclude that PFP programs vary in design and implementation and rarely demonstrate success in raising student achievement or improving teacher retention and recruitment. Effects on teacher motivation appear more promising, provided teachers buy into the program's design and implementation.

Our analysis extends the existing literature in several ways:

- It evaluates a **large-scale, multi-district program** with substantial state investment ($2 billion projected through 2030), offering rare insight into system-wide PFP design.
- It leverages **real-world administrative data**—including VAM, observation, and growth scores—from over 79,000 teachers.

- It integrates **longitudinal performance data** to assess behavioral responses, particularly regarding motivation and sustained performance.
- It operationalizes validity typologies (e.g., concurrent, construct, predictive) in an applied policy context.

Across this literature, evaluations of PFP have concentrated overwhelmingly on outcomes (chiefly effects on student achievement) while the validity of the designation process itself has received far less systematic attention. Whether the metrics used to award performance pay actually identify effective teachers, and whether districts apply them consistently, remains largely unexamined at scale. The present study addresses this gap by evaluating one of the nation's largest state-funded PFP programs not only on its outcomes but on the validity of its design and the fidelity of its implementation. There follows an overview of PFP programs in Texas, a state that has long encouraged merit pay for teachers.

Overview of Texas PFP Programs

Texas has long funded teacher PFP. The Governor's Educator Excellence Grant (GEEG, from 2005) and Texas Educator Excellence Grant (TEEG, 2006–2010) both directed three-quarters of their grants to teachers in economically disadvantaged schools; their principal documented effect was on retention, with turnover falling as award size and number increased, though TEEG showed no achievement gains (Springer et al., 2009).

The District Awards for Teacher Excellence (DATE) program, begun in 2008–09, sharply reduced turnover and produced larger TAKS gains where districts used a select-school rather than district-wide approach, even though DATE schools served more disadvantaged students and had lower passing rates overall (Springer et al., 2010). The Houston Independent School District's ASPIRE program (2006), while it was active, used a subject-level rank-order tournament based solely on students' value-added, with awards up to $7,700; it became a cautionary case, revealing free-riding within larger groups, a narrowing of instruction toward tested subjects, and teacher confusion about how awards were determined (Baratz-Snowden, 2007; Imberman & Lovenheim, 2015).

To develop a program that would test their theory of action, the 86th Texas Legislative Session created a program and provided sustainable funding to support districts in designing systems that incentivize great teaching. TIA funding was built into Texas state law as part of House Bill 3 during the 86th Texas Legislature. It is a Tier 1 allotment through the Foundation School Program (FSP), the system through which the state provides funding to districts. This system, grounded in the Texas Education Code, creates a sustainable funding source for districts implementing TIA. Unlike previous state incentive programs, there is no cap on TIA allotment funds or the number of teachers who may earn a designation.

Districts are asked to create their own systems or procedures for designating teachers at three performance levels: Master (95th percentile); Exemplary (80th percentile); and Recognized (67th percentile).[1] These designations, once achieved, are held for five years. Designations must be

[1] The designation structure described here reflects program rules in effect during our study period (2021–2023). In 2025, House Bill 2 (89th Texas Legislature) expanded TIA funding, increased teacher allotments, created an "Enhanced TIA" track for districts adopting comprehensive strategic-compensation systems, and added a fourth, entry-level "Acknowledged" designation (statewide top 50%) effective 2026–27. These changes postdate our data and do not affect the analyses reported here, but readers should note that the program has since evolved (Texas Education Agency, 2025).

based partly on some measure of student growth and partly on observational evaluation using an approved observation rubric. In a departure from most PFP programs, districts may create their own systems within these parameters, giving more, less, or equal weight to student growth and teacher observation scores. Student growth may be measured by improvement in test scores, value-added model (VAM) scores, attainment of certain learning objectives, or evidence from teacher portfolios. Teacher observations are typically conducted using the Texas Teacher Evaluation and Support System (T-TESS), but a minority of districts prefer approved alternative measures, either developed themselves or selected from the wide variety of possible instruments used elsewhere in the county. The focused teacher behaviors are matched as closely as possible to the targeted items in the T-TESS to ensure comparable scores.

The 86th Legislative Session of the State of Texas designated a Texas university to monitor the quality of implementation by reviewing district-designation data and developing a process to verify its validity. The university team was tasked with validating data submitted by school districts and providing an overall evaluation of the systems that were designed by school districts to designate their teachers as master, exemplary, or recognized teachers. The research team did not work directly with the districts.

It is important to note several things about the designation system process. First, it is up to each school district to design a system to designate its teachers. The system must include measures of student growth and teacher observations, but the specifics of how that is done are left up to the districts. Local control over how to design the designation system was held as a core value of the legislation and as a culture within the state. Once school districts have designed their systems, they submit data to the university team, which evaluates the validity of the data and the correlations within it. Functionally, the review team is evaluating whether school districts are rigorously measuring student growth and teacher observation in a way that holds together statistically. The university team does not make the final decision on whether a school district's proposed designation system is approved, but they do provide data that goes into that decision.

In addition to conducting data validation for the district-submitted designation systems, the university review team has been given other support roles. One of these duties is to provide data and analysis on the TIA program as a whole. As part of that function, the team collects data that then drives future decisions at the agency. Additionally, the team aims to inform the field and participate in the larger knowledge-creation and knowledge-sharing around teacher incentives, teacher pay, and strategic compensation. The university team has no direct contact with the school districts. Reports are, however, reviewed and approved by the TEA before being forwarded to the districts.

The Present Study

The elements shaded in blue in Figure 1 are our focus in this study, one from each characterization of success. We organize these three questions using the contemporary, unified conception of validity, in which validity is a property not of a measure itself but of the inferences and decisions drawn from it, each supported by a distinct line of evidence (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Messick, 1995). We correspondingly ask whether the TIA designation decision is supported by three lines of evidence:

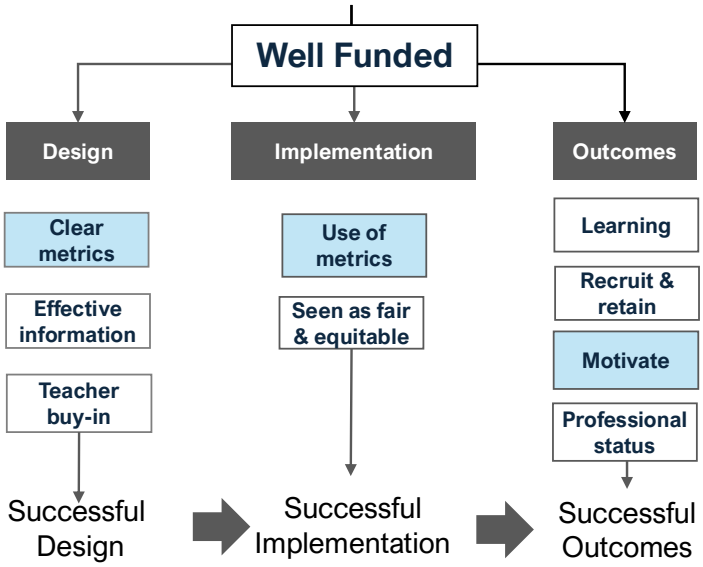
1. **Design Validity** – RQ1: Do designation decisions align with teacher effectiveness indicators? This is a question of *criterion-related (concurrent) validity* — whether designation decisions correspond to concurrent, independent indicators of effectiveness.

2. **Implementation Fidelity** – RQ2: Do districts apply the qualifying metrics consistently and with adequate discriminating power? This is a question of the *reliability and consistency* with which the required measures of student growth and observation are applied across districts.
3. **Outcome Utility** – RQ3: Is designation associated with teacher performance in ways aligned with policy goals? This is a question of *consequential validity*—whether the use of designation is accompanied by the behavioral consequences the policy seeks. Because our design is observational, we treat this as associational evidence about consequences rather than a causal estimate of designation's effect (see Limitations).

This approach contributes to a structured methodology for other jurisdictions evaluating or planning PFP reforms.

Figure 1

Elements of Successful PFP Programs



Method

Sample

The study sample includes 79,198 teachers from 338 local educational agencies (LEAs) (usually referred to as districts) participating in the TIA program from 2021 until the end of 2023. This represented 28% of the 1,202 LEAs in Texas. In the 2026-2027 school year, over 1,000 Texas school districts are projected to participate in this program. It should be noted that school schedules were disrupted during the COVID-19 pandemic in 2020, so we began our analysis for the period after schools reverted to normal schedules, though we acknowledge that residual effects of learning loss from school closures may be present in student achievement data. All teachers were considered designation-eligible by the TEA, and 32% were designated by their LEA for salary bonuses—7% as Master, 13% as Exemplary, and 12% as Recognized. The LEAs vary in size, with an average of 234 teachers per LEA (range = 4-9,451).

Data

In this study, we perform a secondary analysis of data submitted by LEAs as part of the validation checks conducted during the TIA application process. The data include 62,423 teachers from 33 LEAs. About half (46%) of the LEAs were located in rural areas, and about half of the teachers (51%) were teaching courses subject to the State of Texas Assessments of Academic Readiness (STAAR)—Mathematics and Reading Language Arts (RLA) in Grades 3–8, Science in Grades 5 and 7, Social Studies in Grade 8; and Algebra I, Biology, English I & II, and U.S. History in Grades 9–12.

The data contain information about teachers' designation status (Exemplary, Recognized, Master, or no designation), as well as teacher observation scores and student growth scores. We also collected student growth data from the SAS Education Value-Added Assessment System (EVAAS). Both the LEA's and SAS's student growth scores represent the percentage of a teacher's students meeting or exceeding expected growth, and the SAS scores are based on the statewide VAM. VAM is a statistical approach for assessing the effectiveness of schools, LEAs, or teachers by estimating their impact on students' academic growth. It specifically attempts to isolate the educational entity's contribution to learning, controlling for factors such as prior achievement and student demographics. Table 1 presents descriptive statistics for teachers' observation, growth, and VAM scores, reported separately by designation status.

Table 1

Descriptive Statistics for Observation, Growth, and VAM Scores

	Observation Score											
	T-TESS			Other Rubrics (% max scale)			Growth Score			VAM Score		
Designation Level	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Recognized	6,782	3.85	0.38	2,459	0.88	0.09	9,241	0.66	0.16	3,730	0.53	0.16
Exemplary	8,439	4.06	0.40	2,055	0.88	0.10	10,494	0.72	0.15	4,511	0.56	0.16
Master	4,358	4.42	0.44	1,005	0.92	0.09	5,363	0.82	0.15	2,173	0.62	0.17
No designation	42,844	3.37	0.54	11,256	0.76	0.14	54,100	0.51	0.21	21,236	0.45	0.16
All	62,423	3.59	0.61	16,775	0.80	0.14	79,198	0.58	0.22	31,650	0.49	0.17

Note. *M* = mean, *SD* = standard deviation, VAM = value-added model.

Statistical Analysis

Research Question 1: Design Validity – Do designation decisions align with teacher effectiveness indicators? (criterion-related/concurrent validity).

To address the criterion-related/concurrent validity of the LEAs' designation process as one measure of its design success, we examined the extent to which LEAs accurately designate teachers for the PFP incentives. We describe each designation as concordant or discordant with a given performance measure. Discordance can take two forms:

1. A teacher receives a designation even though their score on a given measure falls below the cut-point for the lowest designation level (i.e., Recognized).
2. A teacher does not receive a designation even though their score on a given measure exceeds the threshold for the lowest or a higher designation level.

To achieve designation at any level, the designation system must align with the state's guidance on performance standards. This means, at the lowest level (Recognized), a teacher must satisfy the following criteria: 55% of students meeting or exceeding growth targets, a score of 3.7 on the T-TESS (or 74% on an alternative observation rubric), and/or a VAM score of 55%. The Exemplary level requires 60% growth, a score of 3.9 on the T-TESS (or 78%), and/or a VAM score of 60%, and the Master level requires 70% growth, a score of 4.5 on the T-TESS (or 90%), and/or a VAM score of 70%.

The designations analyzed here are those determined by LEAs under their local designation systems and submitted to the state, where they are checked through the TIA data-validation process. TEA did not necessarily approve these designations; however, since it does not evaluate teachers, it evaluates the system. If that system is found to be overly discordant, it fails data validation. Our discordance indicators are benchmarks of our own construction that compare these submitted designations against individual performance metrics; they do not represent determinations by the agency that a designation was improper.

We used four approaches to assess discordance (designation despite a below-threshold score, or non-designation despite an above-threshold score) for individual teachers: (1) observation score alone, (2) growth score alone, (3) a combination of observation and growth scores, and (4) VAM score alone. The greater the degree of discordance, the less successful this aspect of the program design is. It should be noted that a district incurred a greater penalty in their system's scoring if teachers were designated despite below-threshold scores rather than left undesignated despite above-threshold scores.

We calculated the percentages of each form of discordance at the LEA level and compared them between LEAs that passed data validation checks and those that failed, using bivariate tests. In addition, we conducted generalized linear mixed modeling separately for each teacher-level outcome (yes/no) to examine the effects of teaching assignment (STAAR-tested versus non-STAAR-tested subjects) on the likelihood of each form of discordance. The model was specified as:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1(\text{Assignment})_{ij} + u_j$$

$$u_j \sim N(0, \sigma_u^2),$$

where $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)$ is the log odds (logit) of the binary outcome for teacher *i* nested within LEA *j*; β_0 is the overall intercept expressed as log odds (i.e., the expected log odds of the outcome when Assignment is at its reference category, averaged across LEAs); β_1 is the fixed effect of Assignment,

representing the difference in log odds between STAAR-tested and non-STAAR-tested teachers; u_j is a LEA-level random intercept capturing unexplained variation in the outcome across LEAs, assumed normally distributed with mean zero and variance σ_u^2. The random intercept was estimated using a variance-covariance (VC) structure, which treats LEA-level variance as a single scalar quantity. A logit link function and binomial error distribution were specified, and all fixed effects and variance components were estimated using the Laplace approximation to the marginal likelihood, which provides more accurate parameter estimates than pseudo-likelihood methods for binary outcomes. Fixed effects are reported as odds ratios (e^{β_1}) with 95% confidence intervals.

Research Question 2: Implementation Fidelity – Do districts apply the qualifying metrics consistently and with adequate discriminating power? (reliability and consistency).

This question asks whether school administrators correctly used the two qualifying measures of teacher observation and student growth to make designation decisions. RQ1 examines the accuracy of designation, and RQ2 evaluates the process. We calculated the point-biserial correlation to examine the relationship between teachers' observation and growth scores and their designation status (yes/no), which we used as one measure of implementation success. In addition, we used generalized linear modeling (i.e., logistic regression) to derive areas under the curve (AUC) and Somer's *D* regressing designation status on both observation and growth scores. AUC, a measure of diagnostic accuracy (Swets, 1988), assesses how well the LEAs' chosen metrics distinguish between teachers who received a designation and those who did not. An AUC value of .5 indicates no discriminatory ability, equivalent to chance, while an AUC value of 1 reflects perfect discrimination. Somers' *D*, another measure of classification accuracy, ranges from -1 to +1 (Somers, 1962). Higher positive *D* values suggest stronger alignment between teacher designations and performance measures, indicating effective implementation of the chosen metrics. In contrast, negative *D* values suggest an inverse relationship between designations and performance measures, which could signal potential flaws in the system or underlying data issues. For this calculation, we used data from all previously designated teachers at their highest designation levels, as well as data from newly designated teachers for the current year. We calculated the correlation, AUC, and Somers' *D* within each LEA and compared them across different LEA groups. The weaker the relationship between designation and observation or growth scores, the less successful this aspect of program implementation is.

Research Question 3: Outcome Utility – Is designation associated with teacher performance in ways aligned with policy goals? (consequential validity).

To examine the consequential validity of the LEAs' designation process as represented by motivation, aligned with Vroom's Expectancy Theory, either to maintain or achieve a designation, we inspected the trajectories of 2,404 teachers' observation scores, student growth scores, and VAM scores over a three-year period, from 2021 through 2023. Once again, we should remind readers of the potential negative effects of COVID-19 during the 2021 school year. Teachers fell into three groups:

> *Group A:* Teachers who were never designated by their LEA for salary bonuses (*n* = 684)
> *Group B:* Teachers who received a designation for the first time in 2021 and maintained the same or higher level of designation in subsequent years (*n* = 1,150).
> *Group C:* Teachers who received a designation for the first time in 2022 and maintained the same or higher level of designation in the following year (*n* = 570).

We calculated year-by-year averages of observation, growth, and VAM scores at the LEA level for each group. In addition, we performed general linear mixed modeling separately for each of the three outcomes (observation, growth, and VAM scores) to examine overall group differences (i.e., group effect), changes over time (i.e., year effect), and group differences in these changes (i.e., group-by-year interaction). The model for each outcome was specified as:

$$Y_{ij} = \beta_0 + \beta_1(\text{Group})_{ij} + \beta_2(\text{Year})_{ij} + \beta_3(\text{Group} \times \text{Year})_{ij} + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim N(0, \mathbf{R}),$$

where $Y_{ij}$is the outcome for observation *i* within LEA *j*; $\beta_0$is the intercept; β_1, β_2, and $\beta_3$are fixed effects representing the main effect of group, the main effect of year, and their interaction, respectively; and $\varepsilon_{ij}$is the residual error term. No level-2 random effects were specified; instead, clustering at the LEA level was modeled through the residual covariance structure. Specifically, the repeated measures were defined at the LEA level, and a compound symmetry (CS) covariance structure was imposed on the residuals, which assumes a constant variance across time points and a constant covariance (i.e., equal correlation) between any two observations within the same LEA. This structure accounts for the intraclass correlation arising from teachers being nested within LEAs. All fixed effects were estimated using restricted maximum likelihood (REML), and Type III tests were used to evaluate the statistical significance of each fixed effect. All analyses were conducted using SAS 9.4 (SAS Institute Inc., 2024).

Results

Design Validity

Because districts combine multiple measures using locally determined weights, the operative indicator of design validity is the rate computed when observation and growth are considered together: only 3% of teachers were designated despite below-threshold scores and 8% not designated despite above-threshold scores under this combined criterion. The single-metric rates reported below are diagnostic decompositions, showing how designation would appear if any one measure were applied in isolation, rather than counts of teachers with discordant designations. The 21% designated despite a below-threshold observation score, for example, reflects teachers who fell below the observation cut-point but could legitimately qualify through strong growth, consistent with the program's multi-metric design. Throughout, these rates describe the designations districts submitted under their local systems, not designations separately approved by the agency; the distinction between what a district proposed and what was ultimately approved should be kept in view when interpreting these figures.

The average percentages of each form of discordance are presented in Table 2, separately for (1) all 338 LEAs, (2) the 294 LEAs that passed the data validation checks, and (3) the 44 LEAs that failed validation. Failing validation occurs when the data for a given LEA do not reach the cut-off point in one or more of the checks performed to test the validity of the scores they submit. This table also includes the results of *t*-test and effect sizes for the differences between the LEA groups. Overall, 21% of teachers received a designation through their LEA despite their observation score falling below the standard cut-point for the lowest designation level (i.e., Recognized). A considerably smaller proportion of teachers with a student growth score below the minimum threshold (10%) were designated for salary bonuses. When both observation and growth scores are considered together, the rate of designation despite below-threshold scores decreases to 3%. However, this rate is highest at 25% when only VAM score is considered. In general, this form of discordance is more prevalent among LEAs that failed the data verification process (see Table 2).

Table 2

Designation Accuracy: Discordance Rates, Score Correlations, and Classification Performance at the LEA Level

	All LEAs (*n* = 338)	Passed (*n* = 294)	Failed (*n* = 44)	Passed vs. Failed		
				t	*p*	Cohen's *d*
Designated despite a below-threshold score on						
Observation	**21 ± 24 %**	20 ± 23 %	30 ± 27 %	2.71	.007	0.44
Growth	**10 ± 16 %**	9 ± 14 %	15 ± 24 %	1.71	.094	0.39
Both observation and growth	**3 ± 8 %**	3 ± 7 %	6 ± 13 %	1.57	.124	0.39
VAM	**25 ± 17 %**	24 ± 16 %	27 ± 22 %	0.69	.493	0.14
Not designated despite an above-threshold score on						
Observation	**25 ± 18 %**	25 ± 18 %	26 ± 19 %	0.15	.879	0.02
Growth	**39 ± 24 %**	39 ± 23 %	43 ± 28 %	1.03	.305	0.17
Both observation and growth	**8 ± 11 %**	9 ± 11 %	6 ± 9 %	1.51	.133	0.24
VAM	**10 ± 9 %**	10 ± 9 %	12 ± 12 %	0.75	.454	0.15
r of designation (yes/no) with						
Observation	**.53 ± .16**	.55 ± .15	.41 ± .19	4.36	<.001	0.88
Growth	**.51 ± .17**	.53 ± .16	.39 ± .22	3.77	<.001	0.79
AUC	**.95 ± .05**	.95 ± .04	.91 ± .07	1.85	.073	0.88
Somers' D	**.89 ± .09**	.90 ± .08	.82 ± .14	1.85	.073	0.88

Note. Mean ± standard deviation. LEA = local educational agency, VAM = value-added model, AUC = area under the curve. All figures reflect designations as submitted by LEAs and run through the TIA data-validation process; columns distinguish LEAs that passed from those that failed validation.

The results of generalized linear mixed modeling showed that, as compared to teachers who teach STAAR-tested subjects, those who teach non-STAAR-tested subjects were more likely to be designated despite a below-threshold score when only observation score is considered (odds ratio [OR] = 1.35 [1.24, 1.48]) but they were less likely to be designated despite a below-threshold score when only growth score is considered (OR = 0.67 [0.61, 0.73]). However, no significant difference was found when both growth and observation scores are accounted for. Similarly, teachers with non-STAAR-tested subjects were less likely to be left undesignated despite an above-threshold score when only observation score is considered (OR = 0.79 [0.75, 0.82]) but they were more likely to be left undesignated despite an above-threshold score when either only growth score (OR = 1.42 [1.37, 1.48]) or both observation and growth scores (OR = 1.19 [1.12, 1.27]) are considered. When only VAM score is considered, teaching assignment had significant effects on neither form of discordance. These patterns likely reflect structural differences in how performance data are collected and applied across teaching assignments. Teachers in non-STAAR-tested subjects lack standardized student growth measures, so their growth is captured through locally developed or portfolio-based instruments. For these teachers, designation tracked observation only loosely: i.e., they were more likely to be designated despite below-threshold observation scores and less likely to be passed over when observation was high, while high scores on their locally developed growth measures were rewarded inconsistently, leaving them more likely to be left undesignated despite above-threshold growth scores. This is consistent with locally developed growth measures being applied more variably than standardized assessments. The null finding for VAM is consistent with this interpretation: VAM scores are available only for teachers in tested subjects, rendering teaching assignment irrelevant when that metric is the sole criterion. Notably, combining observation and growth eliminated the assignment-related gap among teachers designated despite below-threshold scores, but not among those left undesignated despite above-threshold scores, with non-STAAR teachers remaining more likely to be overlooked (OR = 1.19). This result could be attributed to many factors, including districts' relatively conservative approaches to teacher designation despite some data that might prompt a more aggressive approach. District decisions can be swayed by both the data and a desire to be careful in their designation systems. Collectively, these results underscore the importance of using multiple, complementary performance measures and suggest that, because non-STAAR teachers are designated at different rates, LEAs may benefit from additional guidance to ensure strong designation processes for teachers across all subject areas.

There are multiple factors to consider in Table 2. First, we see that Failed districts have higher rates of discordance and variance on all levels. Next, there is evidence of a conservative bent in some districts when designating teachers. There may be many reasons for this, but they are not the purview of this paper. Notably, the rate of non-designation despite above-threshold scores was also low, at 10%, when only the VAM score was considered.

These results generally indicate a successful program design with respect to the parameters used to designate merit pay. Of greatest importance is that, when both observation and growth scores are considered, only 3% of teachers were designated despite below-threshold scores (i.e., received a designation that a single measure alone would not support). A larger percentage of teachers (8%) were not designated despite above-threshold scores (i.e., were not designated even though they met the cut-off on a given measure). When growth or observation scores are considered separately, the percentages of discordance are substantially higher, underscoring the importance of considering both parameters when designating teachers for merit pay. Some of these discrepancies may be accounted for by relative differences in the weightings that districts assigned to observation and growth scores, as well as potential differences in observation scores among appraisers who lack calibration.

Implementation Fidelity

Table 2 also reports the average correlations of teachers' designation status (yes/no) with teacher observation and student growth scores, as well as the average classification accuracy as measured by AUC and Somers' *D*. Similarly to RQ1, the results are presented separately for (1) all LEAs (n = 338), (2) the LEAs that passed the data validation checks (n = 294), and (3) the LEAs that failed validation (n = 44). Both observation and growth scores were strongly correlated with designation status (average r = .53 and .51, respectively), suggesting that teachers with higher scores were more likely to be designated by their LEA for salary bonuses. Notably, the correlations were stronger among LEAs that passed the data validation process (see Table 2).

The regression model incorporating both observation and growth scores effectively distinguished designated teachers from non-designated teachers (average AUC = .95, average D = .89). These results demonstrated the effectiveness of the selected metrics in guiding LEA designation decisions. In addition, classification accuracy remained consistently strong across LEA groups, regardless of whether they had successfully completed data validation (see Table 2).

These results suggest that LEAs are appropriately applying the requirements of the program for identifying higher quality teachers for merit pay, demonstrating that the program is largely being successfully implemented. LEA's may choose the relative weights they assign to growth and observation measures and that may account for some of the variation.

Outcome Utility

Figure 2 displays the year-by-year average scores for teacher observation, student growth, and VAM across the three teacher groups. Teachers who were never designated by their LEA (Group A) consistently scored lower than their designated counterparts (Groups B and C), with averages falling below the minimum thresholds for the lowest designation level (55% for growth, 74% for observation, and 55% for VAM).

For teachers who received their first designation in 2021 (Group B), scores remained relatively stable in subsequent years. Small increases are found in observation scores (1-4%) and some decline is registered in VAM scores (2–4%), and growth scores (1-2%).

In contrast, teachers who were not designated in 2021 but received their first designation in 2022 (Group C) experienced significant score increases in the year of their designation. While growth scores declined slightly the following year, they remained above the minimum thresholds. Notably, gains in observation and VAM scores were sustained after the designation year.

Table 3 presents the results from general linear mixed modeling that examined overall group differences (i.e., group effect), changes over time (i.e., time effect), and group differences in these changes (i.e., group-by-time interaction). Significant group-by-time interactions were found for observation and growth scores (both $p < .001$), supporting distinct score trajectories among the three groups. However, the interaction for VAM score was not statistically significant ($p = .15$).

Of the 2,404 teachers in this sample, 684 (28%) were never designated (Group A), 1,150 (48%) were first designated in 2021 and retained at least that level in subsequent years (Group B), and 570 (24%) were first designated in 2022 (Group C). As reported above, the multilevel models indicated distinct observation and growth trajectories across these groups but no significant differences in VAM trajectories. We consider what these patterns may imply for teacher motivation in the Discussion.

Figure 2

Observation, Growth, and VAM Scores in 2021, 2022, and 2023

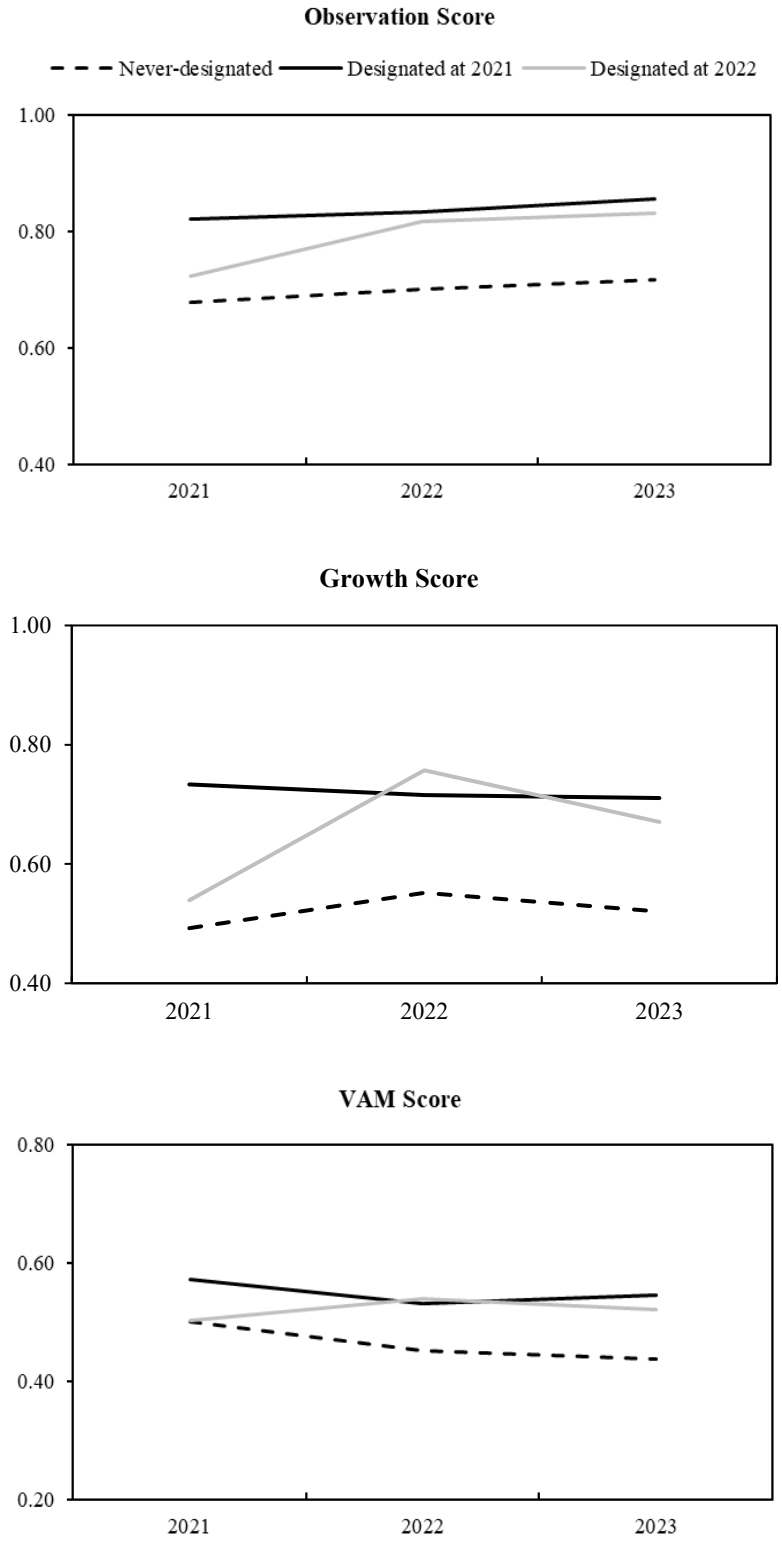


Table 3

Results from Hierarchical Linear Modeling

	Observation Score					Growth Score					VAM Score				
Effect	*b*	*SE*	*F*	*df*	*p*	*b*	*SE*	*F*	*df*	*p*	*b*	*SE*	*F*	*df*	*p*
Intercept	0.83	0.01				0.67	0.02				0.53	0.02			
Group			199.28	2, 72	< .001			130.94	2, 72	< .001			16.84	2, 72	< .001
Never-designated (A)	-0.11	0.01				-0.15	0.02				-0.09	0.03			
Designated at 2021 (B)	0.02	0.01				0.03	0.02				0.02	0.03			
Designated at 2022 (C) (ref.)	0.00	–				0.00	–				0.00	–			
Time			40.78	2, 76	< .001			24.88	2, 76	< .001			1.02	2, 76	.365
2021	-0.11	0.01				-0.13	0.02				-0.02	0.03			
2022	-0.01	0.01				0.09	0.02				0.02	0.03			
2023 (ref.)	0.00	–				0.00	–				0.00	–			
Group × Time			7.75	4, 139	< .001			16.68	4, 139	< .001			1.70	4, 139	.154
A × 2021	0.07	0.02				0.10	0.03				0.08	0.04			
A × 2022	0.00	0.02				-0.06	0.03				-0.01	0.04			
A × 2023 (ref.)	0.00	–				0.00	–				0.00	–			
B × 2021	0.07	0.02				0.16	0.03				0.05	0.04			
B × 2022	0.00	0.02				-0.08	0.03				-0.03	0.04			
B × 2023 (ref.)	0.00	–				0.00	–				0.00	–			
C × 2021 (ref.)	0.00	–				0.00	–				0.00	–			
C × 2022 (ref.)	0.00	–				0.00	–				0.00	–			
C × 2023 (ref.)	0.00	–				0.00	–				0.00	–			

Note. SE = standard error, VAM = value-added model. *F*, *df*, and *p* values were reported from the Type III test of fixed effects.

Policy Relevant Findings and Discussion

TIA Design Validity

Findings show that when combining observation and growth measures, **only 3% of teachers were designated despite below-threshold scores**, and **8% were not designated despite above-threshold scores**, a balance generally favorable in policy terms. This pattern is consistent with both rigor and equity and helps mitigate common concerns that PFP programs either misclassify teachers or reward the wrong behaviors. For policymakers, this suggests that incentive systems can indeed classify performance reliably when multiple indicators are used. The fact that the rate of designation despite below-threshold scores remains so low also helps protect the system's perceived legitimacy—important for sustaining political and institutional support.

Equally important, Texas's decision to allow LEAs flexibility in metric weighting while preserving statewide performance standards appears to contribute to this pattern. This balance between local control and state-level guidance is a compelling model for other states aiming to implement performance-based incentives without incurring backlash over rigidity or irrelevance. In essence, the TIA provides a testable framework for decentralized accountability—an increasingly sought-after feature in modern education policy.

The designation-accuracy analyses also brought to light an equity consideration tied to teaching assignment. Under a single metric, teachers in non-STAAR-tested subjects were more likely to be designated despite a below-threshold observation score, and to go undesignated despite an above-threshold growth score, a pattern consistent with their designations resting on locally developed growth measures that are applied more variably than standardized assessments. Combining observation and growth removed the assignment gap for teachers designated despite below-threshold scores, but non-STAAR teachers remained significantly more likely to be left undesignated despite above-threshold scores even under the combined criterion. The program's multi-metric design, therefore, mitigates some, but not all, assignment-based disparity: non-STAAR teachers continue to face a higher risk of under-recognition. The state has promoted the use of multiple optional components for non-STAAR-tested subjects, including portfolio submissions, testing student learning outcomes through pre- and post-testing, and additional ideas for districts to consider. In the end, this is a district decision. This paper adds to the body of evidence supporting the state policy that using multiple measures for teacher evaluation is a best practice.

TIA Implementation Fidelity

The high correlation between designation decisions and performance measures (AUC = 0.95) suggests that LEAs are largely making data-aligned designation decisions. This is critical from a policy perspective, consistent with districts interpreting and applying state guidance as intended. Additionally, the uniformity of classification accuracy across most districts, even with variation in design, reinforces the strength of the framework.

TIA Outcome Utility (Motivation)

Our data show that the prospect and receipt of designation are associated with improved or sustained performance, most visibly in the year leading up to a teacher's first designation. We interpret this cautiously. Because teachers were not randomly assigned to designation, these associations cannot establish that designation caused the observed performance patterns, and we do not measure motivation directly (see Limitations). With that caveat, the patterns are at least consistent with the possibility that the TIA's structure (its five-year designation window and tiered

status levels) supports sustained engagement rather than the short-lived spikes in effort that annual incentive cycles are often criticized for producing.

Two features of the data are, however, notable for policy. First, some previously non-designated teachers raised their measured performance enough to earn a designation, consistent with the idea that strategic compensation may do more than sort existing talent, though our design cannot distinguish genuine improvement from regression to the mean or pre-existing upward trajectories. Second, teachers designated early did not show subsequent declines in performance, which runs counter to the concern that a guaranteed multi-year bonus erodes effort. We offer the interpretation that designation may carry non-financial value, such as professional recognition or self-efficacy, as a hypothesis for future research rather than a conclusion, since our data cannot lead us to that assumption. If it holds, it would invite policy thinking about how designation status might be linked to leadership roles, mentorship, or professional learning pathways.

Limitations

Several limitations qualify these findings. First, and most fundamentally, teachers were not randomly assigned to designation. Teachers who earned designations (Groups B and C) did so on the basis of prior performance and likely differ from never-designated teachers (Group A) in baseline effectiveness and other characteristics that predate designation. Our models, therefore, identify differences in performance trajectories across self-selected groups; they cannot isolate the effect of designation from the pre-existing differences that led to teachers being designated. A regression discontinuity design around the designation thresholds would offer stronger identification, but its application here is complicated by a defining feature of the program: because districts set their own metric weights and may use different qualifying measures, there is no single, sharp statewide cut score to serve as a running variable.

Second, we do not measure teacher motivation directly. We observe trajectories in observation, growth, and VAM scores and interpret them in motivational terms, but the same patterns are consistent with non-motivational explanations, including regression to the mean, the timing of measurement relative to designation, and ordinary year-to-year variation. Our statements about motivation are accordingly interpretive rather than empirical.

Third, the study period coincides with the disruption and recovery surrounding the COVID-19 pandemic. Although we began our analysis after schools returned to normal schedules, residual learning loss and measurement irregularities may persist in the early years of our data, and the TIA's recent origin limits the availability of a clean pre-pandemic baseline for comparison. Results from these years should be read with that context in mind.

Fourth, our analyses draw in part on value-added measures from EVAAS. VAM estimates are widely documented to be sensitive to model specification, to vary across years for the same teacher, and to be affected by the non-random sorting of students, and their validity for high-stakes use remains contested (American Statistical Association, 2014; Wang et al., 2025). We treat VAM as one external criterion rather than a definitive measure of effectiveness, and we note that the group-by-time interaction for VAM was not statistically significant, which limits inferences about differential VAM trajectories.

Conclusion

The TIA appears to strike a productive balance between policy ambition and design integrity. Unlike many PFP initiatives that collapse under their own weight or lose relevance over

time, TIA offers a promising path forward: one that policymakers in other states and countries may study closely.

This analysis suggests that, under the right conditions, strategic compensation can identify high-performing teachers and is associated with improved and sustained performance over time. With thoughtful attention to design, capacity, and flexibility, policymakers can build durable PFP systems that move beyond pilot projects to meaningful, lasting reform.

From a policy lens, several key takeaways emerge. First, states should not shy away from empowering local education agencies to tailor their designation criteria—provided the system is grounded in shared definitions of rigor and equity. Second, state-level infrastructure to validate and support local implementation is essential; without it, the credibility and sustainability of any merit-pay program may be compromised. Third, longitudinal metrics provide a valuable tool not only for evaluation but also for feedback and continuous improvement, making it possible to detect whether PFP programs incentivize desirable behaviors or introduce unintended distortions.

Finally, TIA highlights the possibility of moving beyond the binary debate of merit pay "working" or "not working." Instead, it invites us to focus on the specific conditions under which performance-based compensation is more or less likely to succeed. These conditions include political will, robust data systems, sufficient funding, and trust between educators and administrators. When these conditions are met, as they appear to be in Texas, strategic compensation can shift from a controversial policy experiment to a sustainable mechanism for professionalizing and energizing the teaching workforce.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. American Statistical Association. https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Azordegan, J., Byrnett, P., Campbell, K., Greenman, J., & Coulter, T. (2005). *Diversifying teacher compensation*. Education Commission of the States.

Baratz-Snowden, J. C. (2007). *The future of teacher compensation: Déjà vu or something new?* Center for American Progress.

Ballou, D., & Podgursky, M. (1993). Teachers' attitudes toward merit pay: Examining conventional wisdom. *ILR Review, 47*(1), 50–61. https://doi.org/10.1177/ 001979399304700104

Barlevy, G., & Neal, D. (2012). Pay for percentile. *American Economic Review, 102*(5), 1805–1831. https://doi.org/10.1257/aer.102.5.1805

Bradford, C., & Braaten, M. (2018). Teacher evaluation and the demoralization of teachers. *Teaching and Teacher Education, 75*, 49–59. https://doi.org/ 10.1016/j.tate.2018.05.017

Breeding, M. E., Béteille, T., & Evans, D. K. (2021). *Teacher pay-for-performance: What works? Where? And how?* World Bank.

Briggs, D., Diaz-Bilello, E., Maul, A., Turner, M., & Bibilos, C. (2014). *Denver ProComp evaluation report: 2010–2012*. Center for Assessment, Design, Research and Evaluation/National Center for the Improvement of Educational Assessment.

Cornett, L. M., & Gaines, G. F. (1994). *Reflecting on ten years of incentive programs: The 1993 SREB Career Ladder Clearinghouse Survey*. Southern Regional Education Board Career Ladder Clearinghouse.

Deneire, A., Vanhoof, J., Faddar, J., Gijbels, D., & Van Petegem, P. (2014). Characteristics of appraisal systems that promote job satisfaction of teachers. *Education Research and Perspectives, 41*, 94–114. https://doi.org/10.70953/erpv41.14005

Farkas, S., Johnson, J., & Duffett, A. (2003). *Stand by me: What teachers really think about unions, merit pay, and other professional matters.* Public Agenda Foundation.

Ford, T. G., & Hewitt, K. (2020). Better integrating summative and formative goals in the design of next generation teacher evaluation systems. *Education Policy Analysis Archives, 28,* 63. https://doi.org/10.14507/epaa.28.5024

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the efficacy of teacher incentives through loss aversion: A field experiment* (NBER Working Paper No. 18237). National Bureau of Economic Research. https://doi.org/10.3386/w18237

Glazerman, S., Chiang, H., Wellington, A., Constantine, J., & Player, D. (2011). *Impacts of performance pay under the Teacher Incentive Fund: Study design report.* Mathematica Policy Research, Inc.

Goldhaber, D., DeArmond, M., & DeBurgomaster, S. (2011). Teacher attitudes about compensation reform: Implications for reform implementation. *ILR Review, 64*(3), 441–463. https://doi.org/10.1177/001979391106400302

Hanushek, E. A., Luo, J., Morgan, A., Nguyen, M., Ost, B., Rivkin, S., & Shakeel, A. (2023). The effects of comprehensive educator evaluation and pay reform on achievement. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.4407482

Heneman III, H. G., & Milanowski, A. T. (1999). Teachers' attitudes about teacher bonuses under school-based performance award programs. *Journal of Personnel Evaluation in Education, 12,* 327–341. https://doi.org/10.1023/a:1008063827691

Hill, A. J., & Jones, D. B. (2020). The impacts of performance pay on teacher effectiveness and retention: Does teacher gender matter? *Journal of Human Resources, 55*(1), 349-385. https://doi.org/10.3368/jhr.55.2.0216.7719r3

Imberman, S. A. (2015). *How effective are financial incentives for teachers?* IZA World of Labor. https://doi.org/10.15185/izawol.158

Imberman, S. A., & Lovenheim, M. F. (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics, 97*(2), 364–386. https://doi.org/10.1162/rest_a_00486

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of public Economics, 89*(5-6), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Jacob, B., & Springer, M. G. (2007). *Teacher attitudes on pay for performance: A pilot study.* (Working Paper 2007-06). National Center on Performance Incentives.

King, K. M., & Paufler, N. A. (2020). Excavating theory in teacher evaluation: Implementing evaluation frameworks as Wengerian boundary objects. *Education Policy Analysis Archives, 28* (57). https://doi.org/10.14507/epaa.28.5020

Kirksey, J. J., Lansford, T., Crevar, A. R., & Mansell, K. E. (2024). *From incentive to impact: The Texas Teacher Incentive Allotment's path to improved retention and achievement.* [White paper]. Center for Innovative Research in Change, Leadership, and Education, Texas Tech University.

Lee, J., Strong, M., Hamman, D., & Zeng, Y. (2021). Measuring teacher buy-in for the Texas pay-for-performance program. *Frontiers in Education, 6*, 729821. https://doi.org/10.3389/feduc.2021.729821

Loeb, S., & Myung, J. (2020). Economic approaches to teacher recruitment and retention. In *The economics of education* (pp. 403–414). Academic Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Milanowski, A. (2003). Varieties of knowledge and skill-based pay design. *Education Policy Analysis Archives, 11*(4). https://doi.org/10.14507/epaa.v11n4.2003

Milanowski, A. (2007). Performance pay system preferences of students preparing to be teachers. *Education Finance and Policy, 2*(2), 111–132. https://doi.org/10.1162/edfp.2007.2.2.111

Murnane, R., & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, *56*(1), 1–18. https://doi.org/10.17763/haer.56.1.l8q2334243271116

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2021). Teacher merit pay: A meta-analysis. *American Educational Research Journal, 58*(3), 527–566. https://doi.org/10.3102/0002831220905580

Reddy, L. A., Dudek, C. M., Peters, S., Alperin, A., Kettler, R. J., & Kurz, A. (2018). Teachers' and school administrators' attitudes and beliefs of teacher evaluation: A preliminary investigation of high poverty school districts. *Educational Assessment, Evaluation and Accountability, 30,* 47–70. https//doi.org/10.1007/s11092-017-9263-3

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2019). *Evaluation: A systematic approach* (8th ed.). SAGE Publications.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

SAS Institute Inc. (2024). *SAS 9.4* [Computer software]. SAS Institute Inc.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review, 27*(6), 799–811. https://doi.org/10.2307/2090408

Springer, M. G., Ballou, D., Hamilton, L., Le, V. N., Lockwood, J., McCaffrey, D. F., Pepper, M., & Stecher, B. (2011). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching (POINT).* National Center on Performance Incentives, Vanderbilt University.

Springer, M. G., Lewis, J. L., Ehlert, M. W., Podgursky, M. J., Crader, G. D., Taylor, L. L., ... & Stuit, D. A. (2010). *District Awards for Teacher Excellence (DATE) Program: Final evaluation report.* National Center on Performance Incentives, Vanderbilt University.

Springer, M. G., Lewis, J. L., Podgursky, M. J., Ehlert, M. W., Gronberg, T. J., Hamilton, L. S., Jansen, D. W., Stecher, B. M., Taylor, L. L., Lopez, O. S., & Peng, A. (2009). *Texas Educator Excellence Grant (TEEG) Program: Year Three evaluation report.* National Center on Performance Incentives, Vanderbilt University.

Springer, M. G., & Taylor, L. L. (2016). Designing incentives for public school teachers: Evidence from a Texas incentive pay program. *Journal of Education Finance, 41*(3), 344–381. https://doi.org/10.1353/jef.2016.0001

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Vroom, V. H. (1964). *Work and motivation.* Wiley.

Wang, Y., Pivovarova, M., & Amrein-Beardsley, A. (2025). Exploring differential effects of a teacher incentive program on student achievement. *Journal of Education, Innovation, and Communication*, *7*(2). https://doi.org/10.34097/jeicom-7-2-2

West, K. L., & Mykerezi, E. (2011). Teachers' unions and compensation: The impact of collective bargaining on salary schedules and performance pay schemes. *Economics of Education Review*, *30*(1), 99–108. https://doi.org/10.1016/j.econedurev.2010.07.007

Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review, 30*(3), 404–418. https://doi.org/10.1016/j.econedurev.2010.12.008

Yuan, K., Le, V. N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis, 35*(1), 3–22. https://doi.org/10.3102/0162373712462625

About the Authors

Jaehoon Lee
Texas Tech University
Jaehoon.Lee@ttu.edu
https://orcid.org/0000-0002-5040-843X
Jaehoon Lee, PhD, is an associate professor of educational psychology at Texas Tech University. His scholarly expertise spans advanced research methodologies, statistical modeling techniques, and measurement instruments applied across education, psychology, public health, and related disciplines. His recent work focuses on the evaluation and application of mixed-effects models, mixture models, Bayesian approaches, and propensity score methods for analyzing complex data sets.

Michael Strong
Texas Tech University
Michael.Strong@ttu.edu
https://orcid.org/0000-0003-2661-4775
Michael Strong, PhD, is a research scientist in the College of Education at Texas Tech University. He is the former Director of Research at the New Teacher Center at the University of California, Santa Cruz, and at the Center on Deafness at the University of California, San Francisco. His current interests center on observing teaching behavior, applying AI to teaching evaluation, and measuring the effectiveness of teacher pay-for-performance programs.

Kristin Mansell
Texas Tech University
Kristin.mansell@ttu.edu
https://orcid.org/0000-0002-2635-2288
Kristin E. Mansell, PhD, is an assistant professor in the College of Education at Texas Tech University. Her research examines how school systems design roles, routines, and instructional structures to strengthen teaching and learning, with a particular focus on strategic staffing, blended and personalized learning, and instructional improvement in rural and high-need contexts. She leads statewide evaluation projects with the Texas Education Agency, including work on the Strategic Operations staffing model, the School Action Fund, and the Teacher Incentive Allotment, and explores teacher workforce dynamics such as retention, mobility, certification pathways, and the impact of professional learning.

Heather Greenhalgh-Spencer
Texas Tech University
Heather.greenhalgh-spencer@ttu.edu
https://orcid.org/0000-0002-9694-6743
Heather Greenhalgh-Spencer, PhD, is a professor in the Department of Curriculum and Instruction at Texas Tech University and Associate Dean of the Graduate School. Her research emerges at the intersection of educational technology, techno-philosophy, pedagogical innovation, strategic staffing, and school-workforce connections. Greenhalgh-Spencer also researches blended/personalized learning (BL/PL) and the ways that BL/PL can create multiple learning pathways and increased opportunities for all students. Her work intersects with issues in STEM education and the STEM (Engineering) pipeline. Dr. Greenhalgh-Spencer has published in multiple international journals of education. She teaches courses on e-learning, BL/PL pedagogies, pedagogical innovation, data literacy, and educational policy and philosophy.

education policy analysis archives

Volume 34 Number 80 September 15, 2026 ISSN 1068-2341