Screen or Barrier? Science of Reading Certification Exams and Teacher Supply in Texas. *

Andrew Avitabile
University of Virginia
Click here for the latest version

August 31, 2026

Abstract

Twenty states now require prospective teachers to pass a standalone Science of Reading exam. This paper provides the first direct evidence on how such a requirement reshapes the teacher pipeline, studying Texas' Science of Reading exam, mandated in 2021 for new elementary and middle-grades candidates in English-related fields. The exam is both a barrier and a screen. As a barrier, it binds at two margins. Candidates who narrowly fail are 9.2 percentage points less likely to earn a Standard certificate within six months and 6.5 percentage points less likely to be teaching the following October, with the employment cost concentrated among alternatively certified candidates, for whom passing gates entry to the classroom. At the market level, new certifications in affected fields surged ahead of implementation as candidates rushed to certify under the old rules, then fell and have remained well below comparison fields within the same preparation programs five years later, a genuine slowdown in the flow of certified teachers rather than substitution across fields. As a screen, the exam carries its information in the score rather than at the cutoff. Scores predict student reading achievement two to four times more strongly than general pedagogy scores and retain that predictive power when both exams are entered together, while teachers just above and just below the passing threshold have indistinguishable value-added in reading even after correcting for selection into teaching. Extending the score–value-added relationship below the cutoff nonetheless implies that a lower threshold would admit teachers who are, on average, somewhat less effective. Early evidence from Texas Reading Academies, state-supported structured coursework in the Science of Reading, is consistent with states helping preparation programs raise pass rates, though this evidence is preliminary.

*Andrew Avitabile: yaj3ma@virginia.edu. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200005 to the University of Virginia. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education. I thank seminar participants at the University of Virginia's EdPolicyWorks PolicyLab for helpful comments. Any remaining errors are my own.

1 Introduction

Teacher certification exams occupy an uneasy position in education policy. Licensure exams are a *barrier* to teacher supply (Angrist & Guryan, 2008; Goldhaber, 2007; M. Kleiner, 2006). Yet weakening requirements carries its own risks, since exams can be valuable *screens* insofar as they filter candidates on knowledge relevant to student learning (Stiglitz, 1975). Whether an exam is worth its supply costs thus depends on how much screening it actually does. Consistent with screening, teachers who entered through relaxed pandemic-era pathways were less effective than their traditionally certified peers (Backes et al., 2024; J. J. Kirksey, 2024). At the same time, there is limited evidence that raising passing thresholds on certification exams improves the quality of teachers who clear them (Deneault et al., 2025; Orellana & Winters, 2025). Faced with this tension, some states have eliminated certification exams in response to teacher shortages and equity concerns (Swisher, 2022), while others have expanded testing requirements (National Council on Teacher Quality, 2024).

That expansion is most pronounced in early literacy. Decades of research show that effective reading instruction requires explicit, systematic teaching of foundational skills such as phonological awareness, phonics, and fluency (Castles et al., 2018; Seidenberg, 2017), but licensure exams have historically failed to measure whether teachers understand these practices (Ellis et al., 2023). In response, 20 states now require a standalone Science of Teaching Reading (SOR) exam as a condition of certification (National Council on Teacher Quality, 2024), betting on evidence that comprehensive SOR reforms improve student reading outcomes (Novicoff & Dee, 2025; Spencer, 2024). The extent to which SOR exams are screens, or only raise barriers, remains an open question.

This paper provides the first direct evidence on that question. I study the Texas Examinations of Educator Standards (TExES) SOR exam, announced in June 2019 as part of House Bill 3 and required beginning January 1, 2021 of all new candidates in early childhood through grade 6, core subjects grades 4–8, and English language arts and reading grades 4–8. Texas layered this exam onto an existing certification pipeline rather than replacing a test or moving a cutoff. The requirement therefore added a new hurdle—registration and retake fees, preparation time, and a thirty-day wait between attempts—to fields accounting

for close to two-fifths of the state's newly issued teaching certificates, making the rollout a natural setting in which to weigh an exam's costs against what it measures.

The tension between screen and barrier motivates three research questions:

1. Does the SOR requirement act as a *barrier*, delaying certification and reducing the flow of newly certified teachers?

2. Does the exam *screen*, with scores that identify candidates more effective at raising student achievement?

3. If the exam screens on real knowledge, can structured preparation help more candidates meet the standard rather than lowering it?

To answer these questions, I combine administrative data on the universe of Texas certification exams, certificates, and public school employment, linked to student achievement records. I explore the effect of the exam as a barrier in two ways. A regression discontinuity at the passing threshold identifies the effect of narrowly failing, and a difference-in-differences design compares the flow of new certifications in fields that began requiring the exam to fields within the same preparation programs that did not. I test for screening on three margins. First, I use value-added models that relate teachers' SOR scores to their students' achievement. I then estimate a regression discontinuity testing whether teacher effectiveness jumps at the passing threshold. Thirdly, I extrapolate the estimated relationship between exam score and teacher value-added to explore how effective the marginal teacher admitted under a lower threshold would be. Finally, I ask whether Texas Reading Academies (TRAs), state-sponsored structured coursework aligned to the exam and delivered by educator preparation programs (EPPs), can help candidates clear the bar.

I find that the exam is both a barrier and a screen. As a barrier, narrowly failing reduces the probability of earning a Standard certificate within six months by 9.2 percentage points and of teaching the following October by 6.5 percentage points. These costs are not evenly distributed. They fall hardest on alternatively certified candidates, for whom passing the exam gates entry to the paid internship that puts them in a classroom.[1] A candidate on

[1] As per 19 Tex. Admin. Code § 230.36, an alternatively certified candidate must pass all required content certification examinations before an EPP can recommend the candidate for an intern certificate.

that route who narrowly fails is 12 percentage points less likely to be teaching the following year. For traditional candidates, who sit for the exam at the end of their preparation, the same failure is a delay rather than a derailment.

The market-level results follow the same logic at a larger scale. New certifications in fields where the requirement took effect surged ahead of implementation, as candidates rushed to certify under the old rules, then fell and have remained well below comparison fields within the same EPP five years later. The decline holds under randomization inference appropriate to the small number of underlying certification fields, and under a linear continuation of the pre-period trend into the post-implementation window. It also does not appear to reflect substitution across fields, on either of two margins. Candidates who narrowly fail are no more likely to earn a certificate in fields where the exam is not required, and post-mandate cohorts certify more slowly in affected fields without certifying any faster elsewhere. The contraction is concentrated in the non-traditional routes—alternative, out of state, and by exam—which likely reflects a difference in the marginal cost of the requirement across routes.

The exam screens, but it does so through the score and not the pass/fail threshold. SOR scores predict student reading achievement two to four times more strongly than general pedagogy exam scores, and they retain their predictive power when both exams are entered together in a single regression. This is notable because teacher effectiveness is notoriously difficult to predict from anything observable before hire (Staiger & Rockoff, 2010), and because the broader literature relating certification scores to teacher value-added has reached mixed conclusions (Clotfelter et al., 2007; Cowan et al., 2023; T. Kane & Staiger, 2008; Rockoff et al., 2011). The contrast between the two exams suggests that what an exam measures determines the extent to which it carries signal about teaching ability, and that a test built around specific, well-defined instructional content predicts effectiveness where a test of general pedagogy does not.

The threshold, by contrast, does not sort. Teachers who barely pass and barely fail have similar value-added in reading, and this null survives correcting for the selection of candidates who fail and persevere into the classroom. The corresponding result in math is less conclusive, since the same correction admits a wide range of plausible effects. This adds to a small literature testing whether teacher licensure thresholds screen on quality (Deneault

et al., 2025; Orellana & Winters, 2025). That the threshold fails to sort at its current location does not mean it could be lowered freely. Extending the estimated relationship between score and reading value-added below the cutoff implies lowering the bar by 60 points would admit a marginal teacher 0.026 standard deviations less effective. Given that the exam's predictive power lies in the score rather than the line, policymakers might consider using scores to target support rather than spending them on a single pass-or-fail decision.

Finally, this paper asks whether better preparation can help more candidates clear the bar. Beginning in 2024–25, select EPPs began offering TRAs. Programs in the first adopting cohort saw first-attempt pass rates rise about 4 percentage points relative to comparison programs. I treat this as early and suggestive rather than settled, because the adopting programs are few and because the identifying assumptions are supported for the first cohort and not the second. Read with that caution, the result speaks to a question that is both policy-relevant and almost entirely unstudied. States adding licensure exam requirements want to guarantee minimum competencies, but what states and EPPs can do to prepare candidates to meet them is nearly unexamined, even as the exams demonstrably shape who enters teaching.

The paper makes three contributions. First, it provides the first direct evidence on how adding a Science of Reading exam reshapes the teacher pipeline, a requirement now in force in twenty states and spreading. That the exam's cost depends on where it falls in a candidate's route, and lands heaviest on non-traditional routes, adds a dimension to a literature on licensure and teacher supply that has documented heterogeneity primarily by race. Second, it separates what an exam's score measures from what its threshold screens. SOR scores carry information about teacher effectiveness that general pedagogy exams do not, yet the pass/fail cutoff does not appear to separate more effective candidates from less effective ones, which implies the exam may be better suited to informing preparation than to gatekeeping entry. Third, the paper moves past the binary of whether to test toward the largely unstudied question of how to prepare candidates to meet new certification standards. The evidence here is early, but the question will only grow more pressing as interest in the teaching profession continues to erode (Bartanen & Kwok, 2023; Kraft & Lyon, 2024) and if similar exam requirements expand to fields that already struggle to fill positions (Cowan

et al., 2016), as the nascent "science of teaching math" movement suggests they might (Hankins, 2026).

2 Background

2.1 Conceptual Framework: Licensure as Screen and Barrier

This paper draws on three related economic traditions. The economics of occupational licensing (M. Kleiner, 2006; M. M. Kleiner & Krueger, 2013) treats licensure as a labor market institution performing two functions simultaneously. First, a *screening* function that filters candidates on minimum competency and second a *barrier* function that imposes costs (e.g., fees, preparation time, retake delays, and foregone earnings) on prospective entrants. The welfare consequences of a licensure requirement depend on the balance between these two. If an exam screens effectively, its supply costs may be offset by quality gains. If it imposes barriers without meaningful screening, it restricts supply without compensating benefits.

Whether an exam screens effectively depends on what its scores measure. The classical distinction between *screening* (Stiglitz, 1975) and *signaling* (Spence, 1973) is useful here. A screen identifies real differences in candidate knowledge relevant to performance, while a pure signal produces a credential whose value derives from the cost of obtaining it rather than what it certifies. The distinction has direct implications for policy. If the SOR exam screens on knowledge that matters for student learning, then structured preparation that builds this knowledge (i.e., an investment in human capital (Becker, 1964)) is the appropriate response to supply concerns. If the exam is closer to a pure signal, EPPs may help candidates pass without making them better teachers.

Together, these traditions generate three testable predictions that organize the empirical work. First, if the exam functions as a binding barrier, failing should delay certification and reduce subsequent employment, with larger effects for candidates already at the margin of the labor market. Second, if the exam screens on real knowledge, scores should predict student reading achievement. Third, if barriers stem from inadequate preparation rather than

fixed differences in candidate ability, structured preparation aligned to exam content should improve performance. The remainder of this section reviews the evidence base motivating SOR exams; the literature on certification exams, teacher supply, and quality; and the Texas institutional context in which I test these predictions.

2.2 The Science of Reading and Teacher Preparation

The Science of Reading synthesizes decades of empirical research showing that reading acquisition requires explicit, systematic instruction in phonological awareness, phonics, fluency, vocabulary, and comprehension (Castles et al., 2018; Seidenberg, 2017). Meta-analyses consistently demonstrate the effectiveness of systematic phonics instruction for improving literacy outcomes (Ehri et al., 2001; McArthur et al., 2018). Despite this evidence, teacher preparation programs have historically emphasized approaches (e.g., balanced literacy and whole language) that downplay these methods, leaving classroom literacy instruction without an evidence-based foundation (Moats, 2020). Ellis et al. (2023) analyzed 702 elementary educator preparation programs across all fifty states and found that 30% provide no practice opportunities connected to the core components of reading. Since teacher knowledge of foundational literacy skills predicts student reading growth (Spear-Swerling & Brucker, 2004), and early, intensive, evidence-based instruction can bring most at-risk students to grade-level word-reading ability (Torgesen, 2004), inadequate preparation represents a primary barrier to equitable reading outcomes.

States have responded to this evidence aggressively. As of January 2026, 20 states require a standalone certification exam demonstrating minimum competency in the Science of Teaching Reading (National Council on Teacher Quality, 2024). In some cases, SOR reforms have coincided with improved student outcomes. For example, quasi-experimental evidence from Mississippi's comprehensive literacy reforms indicates gains of 0.23 standard deviations in reading (Spencer, 2024), and Novicoff and Dee (2025) find similarly positive effects of California's Early Literacy Support Block Grant. However, these initiatives bundle curriculum adoption, coaching, professional development, and certification requirements, making it difficult to isolate the contribution of licensure exams specifically. To date, no studies have directly examined whether Science of Reading certification exams improve student outcomes

or how they shape the teacher workforce.

2.3 Teacher Certification Exams, Quality, and Supply

A large literature shows that certification requirements function as binding barriers to entry, shaping both the size and composition of the teacher workforce (Angrist & Guryan, 2008; Boyd et al., 2007; Deneault et al., 2025; Goldhaber, 2007). Angrist and Guryan (2008) demonstrate that more stringent testing requirements reduce teacher supply and raise teacher wages—consistent with restricted entry—without measurable improvements in teacher quality. Most relevant to this paper, Deneault et al. (2025) use Texas certification exam data to document that narrowly failing a certification exam substantially delays entry into teaching and imposes significant earnings losses, with considerably larger effects for underrepresented minority (URM) candidates than for white candidates. These supply costs are especially salient at a moment of declining interest in the teaching profession (Bartanen & Kwok, 2023; Kraft & Lyon, 2024) and persistent gaps between the demographic composition of the teacher workforce and the students it serves.

Whether these barriers are justified depends on the exams' screening value, and here the evidence is less favorable. Teacher effectiveness has proven difficult to predict from anything observable at hire, with credentials and test scores exhibiting at most modest and inconsistent correlations and classroom experience among the few consistent predictors (Clotfelter et al., 2007; T. J. Kane et al., 2008; Rockoff et al., 2011). Some certification scores nonetheless carry predictive content (Cowan et al., 2023), but even where the score predicts, the passing *threshold*—the actual policy instrument—appears to screen poorly. Deneault et al. (2025) also speak to screening and exploit a reform that made Texas elementary certification exams more difficult. Here they find nearly identical value-added distributions among teachers who passed the easier and harder versions, so the harder exam reduced URM supply without improving quality. Orellana and Winters (2025) directly test whether licensure thresholds screen on quality using Connecticut data. Failing the Praxis II reduces the likelihood of certification by about 6.6 percentage points, with deterrent effects concentrated in shortage fields, and rather than weeding out weaker candidates, failing disproportionately deters candidates at least as effective as those who passed. In sum, existing exams appear to

carry limited screening information in their scores, and their thresholds function primarily as barriers.

Despite this evidence, most states continue to require and expand certification testing, for two understandable reasons. First, as preparation pathways grow more heterogeneous with the expansion of alternative certification (National Council on Teacher Quality, 2025), exams are one of the few mechanisms states retain to verify minimum competency. This concern is validated by evidence that teachers entering through relaxed pandemic-era emergency pathways were less effective than traditionally prepared peers (Backes et al., 2024; J. Kirksey & Gottlieb, 2023). Second, as states adopt new instructional approaches such as Science of Reading-based literacy instruction, new exams are created to verify these specific skills.

On the other hand, other states have considered this evidence and moved in the opposite direction. For example, New Jersey eliminated its Praxis Core requirement effective January 2025, citing shortages and financial barriers (NJ State Policy Lab, 2025). Similarly, several states have dropped or made optional the edTPA (Amin, 2022)—a performance-based, subject-specific assessment in which teacher candidates assemble a portfolio during their student teaching, including lesson plans, video of their own instruction, and written commentaries, scored externally to gauge classroom readiness.

Much of this policy debate treats the options as binary. States can maintain stringent requirements that restrict supply and diversity, or they can weaken standards and risk admitting underprepared teachers. This study examines a third approach, shifting the question from *whether to test* to *how to prepare candidates to meet the standard*.

2.4 Texas' Science of Reading Exam

In June 2019, Texas passed House Bill 3, which mandated that new teacher candidates demonstrate proficiency in the Science of Reading through a certification exam. Beginning January 1, 2021, candidates pursuing certifications in early childhood through grade 6, core subjects grades 4–8, or English Language Arts grades 4–8 must pass the TExES Science of Reading exam (exam 293). The exam assesses knowledge of phonological awareness, phonics, fluency, vocabulary, and reading comprehension instruction grounded in evidence-based practices; example items are provided in Appendix B.

The exam consists of 90 multiple-choice questions scored on a scaled system with a passing score of 240. Candidates who fail may retake the exam after a 30-day waiting period, are limited to five attempts absent a waiver, and pay a fee of $116 per attempt, creating direct financial costs for candidates who must retake. The affected certifications account for 41.7% of newly issued teaching certificates in Texas in the pre-mandate period (Table 2), making the SOR exam a significant checkpoint in the teacher certification pipeline.

2.5 Texas Reading Academies at EPPs

To support preservice teachers' preparation in evidence-based early literacy instruction, the TEA authorized select EPPs to offer Texas Reading Academy coursework beginning in the 2024–25 academic year. These academies extend the TRA, a statewide professional development program rooted in the Science of Teaching Reading, to preservice preparation.

TRAs provide a structured, intensive sequence of coursework beyond standard EPP requirements. The curriculum consists of twelve modules covering the full scope of research-based literacy instruction, including oral language and vocabulary development, phonological and phonemic awareness, print concepts and alphabet knowledge, decoding and encoding, reading fluency, reading comprehension, and written composition. Instruction emphasizes explicit, systematic, and data-driven practices, with attention to culturally sustaining pedagogy, emergent bilingual learners, and special populations. Candidates complete online modules delivered through Canvas, participate in in-person instructional sessions, and demonstrate mastery through graded artifacts, assessments, and coaching, with a minimum mastery threshold of 80 percent across required assignments.

EPPs must apply through a TEA-administered authorization process requiring an Authorized Provider within the EPP, Cohort Leaders who pass a multi-stage screening, and required training and onboarding. Once approved, EPPs retain flexibility in integration: some offer academies as intensive standalone courses, while others embed the content across multiple semesters of coursework.

Importantly, participation is not universal. EPPs self-select into the application process, not all applicants are approved, and adoption occurred at different points in time. Table A8 documents the authorized EPPs and each program's adoption year based on TEA adminis-

trative records.[2] This staggered and incomplete adoption generates variation in preservice teachers' access to academy coursework across programs and cohorts, which I exploit in the empirical analysis.

3 Data and Sample

This analysis uses administrative data from the University of Houston Education Research Center (ERC), which links records across multiple state agencies. The ERC integrates data from the TEA, the Texas Higher Education Coordinating Board (THECB), and other state entities through common identifiers, enabling longitudinal tracking of individuals across the K–12, postsecondary, and teacher workforce systems. I draw on four categories of data: teacher certification exams, teacher certifications, public school employment and student achievement, and postsecondary enrollment records.

3.1 Teacher Certification Exams and Certifications

The State Board for Educator Certification (SBEC) maintains two cumulative administrative files that form the backbone of this study. The first is a record of all attempts at SBEC certification exams, covering every exam administered through March 2026. Each record includes the exam type, administration date, scaled score, pass/fail result, certification route (traditional, alternative, or certification by exam), and the EPP that approved the candidate to test. The second file contains all teaching certificates issued in Texas through early 2026. Each certification record includes the certification field, type, and program; the issue and effective dates; the EPP attended; and candidate gender and race/ethnicity. Both files contain unique identifiers that allow records to be linked across files and to other ERC data sources.

The exam file lacks demographics, so I construct race and ethnicity measures for exam takers by linking to two external sources: TEA's Public Education Information Management System (PEIMS) student records, covering individuals who attended Texas public schools as K–12 students, and THECB postsecondary records, covering individuals who applied to

[2]More information on participating EPPs can be found here.

or attended Texas public colleges and universities. Together, these linkages yield race and ethnicity for approximately 65 percent of first-attempt exam takers. The remainder likely attended K–12 school outside Texas and private or out-of-state postsecondary institutions. Analyses requiring demographics restrict to candidates with available covariates, and I note this restriction where it applies.

3.2 Public School Employment and Student Achievement

Employment data come from the PEIMS staff files, annual snapshots of all individuals employed in Texas public schools measured as of the last Friday in October. I observe employment status, district and campus assignments, role classifications, and experience at a single point each year, and I link these records to the exam and certification files to measure whether exam takers work as teachers in the October following their exam. For analyses linking teachers to students, I use TEA's class roster data, which connect teachers directly to the students in their classrooms. Student achievement is measured by State of Texas Assessments of Academic Readiness (STAAR) scores in reading and math, standardized to mean zero and standard deviation one within grade, year, and subject. Student covariates including prior achievement, race, gender, special education status, gifted status, and economic disadvantage, come from the PEIMS student files.

3.3 First-Attempt Certification Exam Takers

The first analytic sample consists of all individuals who sit for the TExES SOR exam for the first time with a valid scaled score, restricted to candidates whose exam date falls at least one year before the end of the certification data.[3] This ensures a minimum of one year of follow-up for certification and employment outcomes. The resulting sample includes 21,692 first-attempt SOR exam takers from January 1, 2021 through October 23, 2023.

Table 1 presents summary statistics overall and by first-attempt pass status. Approximately 80 percent of candidates pass on their first attempt, with a mean scaled score of 251.6 against a passing score of 240. Candidates who fail are disproportionately on the alternative

[3]The scaled-score restriction also excludes the exam's initial piloting period, during which SBEC reported only pass/fail results, so the effective sample begins in September 2021 rather than January 2021.

certification route (52.7 percent, versus 36.3 percent of those who pass), more likely to be non-white (53.0 versus 40.2 percent among the subsample with demographics), and slightly less likely to be female (89.0 versus 92.9 percent).

A notable feature of the data is that approximately 31 percent of first-attempt SOR exam takers are already employed as teachers in Texas public schools before their exam date, at nearly identical rates among those who pass (31.6 percent) and fail (31.1 percent). This reflects the structure of the Texas teacher labor market, in which individuals can teach on interim credentials, emergency permits, or without formal certification in districts of innovation.[4] Accordingly, 58.8 percent of candidates earn a Standard certificate within one year while 70.9 percent are employed as teachers the following October. This is consistent with a context where many teachers enter classrooms before completing all certification requirements.

3.4 Newly Issued Certifications

The second analytic sample consists of all newly issued standalone teaching certificates in Texas from 2016 through 2025, where each observation is the first time an individual receives a particular certification (i.e., individuals who earn multiple standalone certifications may appear multiple times in the dataset).[5] Here, I observe the certification field, type, issue date, EPP, and candidate demographics for all 342,816 records. I classify each certificate as SOR-required or comparison following the Texas Education Agency Required Test Chart. Table A2 lists the full classification.[6]

Table 2 presents summary statistics for this sample, split by whether the certification was issued before or after the SOR mandate's June 2019 announcement. The two periods are broadly similar in route composition, though out-of-state certifications rise modestly, from

[4]These are districts who are allowed by TEA to opt out of regulations on inputs like teacher certification and class sizes. See Anglin (2024) for more information.

[5]I begin the sample in 2016 to avoid the January 2015 replacement of the Generalist EC-6 (TExES 191) and Generalist 4–8 (TExES 111) exams with the Core Subjects EC-6 (TExES 291) and 4–8 (TExES 211) exams.

[6]A certification is coded SOR-required if the TExES Science of Teaching Reading exam (293) is among its required content-pedagogy tests. Non-teaching credentials and supplemental certificates that attach to a base certificate rather than certifying a teaching field independently are excluded from the sample. For more information, see the notes to Table A2.

9.9 to 13.6 percent. Descriptively, the share of newly issued certifications in SOR-required fields fell from 41.7 percent pre-announcement to 36.0 percent afterward. The demographic composition also shifted, with the White share declining from 59.2 to 55.2 percent and the Hispanic share rising from 25.1 to 27.6 percent.

To estimate how the SOR requirement affected the flow of new certifications, I collapse this sample to counts by educator preparation program (EPP) × program-type (e.g., traditional, alternative, etc.) × certification-field × half-year (i.e., January to June and July to December). This collapse spans 200 EPP-program-type units, 3 SOR-required and 16 comparison certification fields, and 2016 January to June (H1) through 2025 July to December (H2), yielding 9,585 SOR-required and 51,120 comparison cells. Table 3 describes the resulting panel.[7]

3.5 Texas Reading Academy Data

For the analysis of Texas Reading Academies, I obtained the set of EPPs authorized to offer academy coursework, and each program's adoption year, from TEA administrative records (Table A8). I match this information to EPP identifiers in the exam and certification files to classify candidates by whether their EPP offered a TRA at the time of their first exam attempt.

Table 4 compares pre-period characteristics of the 10 EPPs that adopted academies in 2024–25 and the 98 EPPs that never adopted, measured over the 2021–22 through 2023–24 academic years. EPPs adopting in 2025–26 are excluded from both groups. Most importantly, the two groups perform similarly on both the SOR and Pedagogy and Professional Responsibilities (PPR) exams in the pre-period. Pass rates and scaled scores are statistically indistinguishable for both exams. The groups are also balanced on certification production, demographics, and the share of certifications in SOR-required fields. The most notable com-

[7] I code a cell as zero if the EPP-program-type issued no certifications in that field during that half-year but was otherwise active, defined as having issued a certification in *any* field during that half-year. Periods before an EPP-program-type's first observed activity, or after its last, are excluded from the panel rather than coded as zero. This distinguishes a program that was operating but chose not to issue a given certification from one that had not yet begun, or had ceased, issuing certifications altogether. Under this rule, 51.2 percent of SOR-required cells and 69.9 percent of comparison cells are zero, reflecting the narrowness of many comparison fields (e.g., Foreign Language, Computer Science) relative to the broader SOR-required fields.

positional difference is that academy EPPs are more likely to be traditional university-based programs (73.3 versus 55.8 percent).

4 Empirical Strategy

4.1 Barrier: The Cost of the Requirement

4.1.1 Individual-Level Cost of Failing

Candidates scoring 240 or above pass the SOR exam and those below fail. Assuming candidates just above and just below this cutoff are comparable in all respects other than pass status, comparing their outcomes identifies the causal effect of passing the exam for candidates at the margin.

Using a regression discontinuity at the passing threshold, I examine four outcomes that trace the consequences of failing from immediate response to career trajectory: whether candidates retake the SOR exam within one year, whether they earn a Standard teaching certificate within six months and within one year of their exam date, and whether they work as a teacher in a Texas public school the following October.[8]

Formally, I estimate

$$Y_i = \alpha + \beta \cdot \mathbb{1}(\text{Score}_i \geq 240) + f(\text{Score}_i - 240) + \varepsilon_i, \tag{1}$$

where $f(\cdot)$ is implemented as a local linear regression with a triangular kernel and the data-driven bandwidth of Calonico et al. (2014). Since employment is observed only annually while exam dates vary throughout the year, exam timing mechanically determines how long candidates have to secure a position before the October snapshot. For consistency, all specifications therefore include month-of-exam fixed effects. As a robustness check, I assess robustness to alternative bandwidth selectors. Since certification routes differ in candidates' institutional position at the time of the exam, with some finishing university-based programs

[8] Teachers are observed in Texas public schools once annually. For example, for a candidate who takes the SOR exam between October 1, 2023, and September 30, 2024, employment is measured as of the October 2024 staffing snapshot for the 2024–25 school year.

and others already teaching on interim credentials, I estimate the model separately by route.

Identification requires that candidates just above and below the cutoff are comparable in all respects other than their pass status, so that any difference in outcomes can be attributed to passing rather than other differences between the groups. The main threat to this assumption is manipulation. If candidates could precisely control their scores around the threshold, those who landed just above might differ systematically from those just below. This is implausible by design, as test scores are computed from raw scores through a conversion candidates do not know in advance, candidates cannot observe their standing relative to the threshold while testing, and no mechanism exists to revise a score once submitted. Two pieces of evidence support the assumption. The score density rises smoothly through the threshold with no bunching on either side (Figure A2), and a formal manipulation test (Cattaneo et al., 2018) fails to reject the null of no sorting at conventional levels (p=0.051). In addition, pre-test covariates are balanced across the threshold (Table A1). These results all indicate that candidates are not sorting across the cutoff.

4.1.2 Market-Level Effects on the Flow of New Certifications

The RD identifies the cost of failing for candidates near the threshold, but cannot speak to the margins where the mandate's barrier may matter most. Candidates who score well below the cutoff on their first attempt face more uncertain paths to passing, and the exam's direct and indirect costs (e.g., fees, study time, travel to a testing center) may deter prospective candidates from the pipeline altogether. To explore this margin, I compare the number of newly issued, standalone teaching certifications in SOR-required fields to the number in comparison fields, using the sample described in Section 3.4.

Using counts of newly issued certifications at the educator preparation program (EPP) by program type (e.g., Traditional, Alternative, etc.) by certification type (e.g., EC-6 generalist) by half-year level (e.g., January to June), I estimate event study regressions via Poisson pseudo-maximum likelihood:

$$\mathbb{E}\left[Y_{ekct}\right] = \exp\left(\sum_{t \neq 2019\text{H1}} \beta_t \cdot \mathbb{1}(t) \times \text{SOR}_c + \gamma_{ek} + \delta_c + \lambda_t\right), \tag{2}$$

where Y_{ekct} is the number of new certifications issued by EPP e through program type k (e.g., traditional, alternative) in certification area c during half-year t, and SOR_c indicates that certification area c started requiring the SOR exam in January 2021. The specification includes EPP-by-program-type (γ_{ek}), certification type (δ_c), and half-year (λ_t) fixed effects, so identification comes from within-EPP-program-type comparisons of how SOR-required and non-SOR certifications evolve over time. The coefficients of interest are the β_t, whose exponents give the rate ratio of certifications in SOR-required relative to non-SOR fields in each half-year, normalized to the first half of 2019. I choose the first half of 2019 as the reference period, as it is the last period before the enactment of House Bill 3 in June 2019, which signed into law that the SOR exam would become a requirement on January 1, 2021. Poisson pseudo-maximum likelihood estimation accommodates the count nature of the outcome and allows estimation with cells containing zero certifications (Santos Silva & Tenreyro, 2006). Standard errors are clustered at the certification type level, the level at which the SOR requirement is assigned. Similar to the RD analyses, I also estimate models separately by certification route.

I complement the event study with a difference-in-differences specification that replaces the half-year interactions with a single post-first half of 2019. This specification estimates the average effect of the policy on certification flows post-announcement, including any behavioral response that occurred in anticipation of the requirement's January 2021 implementation. Once candidates learned of the mandate in June 2019, they had an immediate incentive to certify before the exam became a requirement. Restricting attention to post-2021 responses would ignore this anticipatory behavior and understate the requirement's full effect on the flow of entrants into the profession. I also report a donut specification that excludes half-years 2019H2 through 2021H2 from the estimation sample entirely, removing any anticipatory surge in certifications ahead of implementation and its immediate aftermath rather than merely redefining the post-period indicator. I interpret this specification as the requirement's long-term effect, once the market has settled into its new steady state, and report both throughout the results.

With only 19 certification-type clusters, 3 of which are SOR-required, the asymptotic justification for cluster-robust standard errors is weak. With few clusters—especially in the

treated group—clustered standard errors may understate the true sampling variability of the treatment coefficient (Cameron & Miller, 2015). To correct for this, I assess inference on the difference-in-differences estimates via randomization inference, following MacKinnon and Webb (2020). Specifically, I re-estimate the specification 2,000 times, each time randomly assigning SOR-required status to 3 of the 19 certification fields while holding the true number of treated fields fixed, and compute the t-statistic on the coefficient of interest in each placebo estimate; the reported p-value is the share of these placebo t-statistics whose magnitude equals or exceeds that of the t-statistic from the true assignment. I use the t-statistic rather than the raw coefficient because the certification fields vary considerably in size, and randomization inference based on coefficients alone can be unreliable when clusters are this heterogeneous (MacKinnon & Webb, 2020).[9] Randomization inference does not rely on asymptotic approximations across clusters and remains valid regardless of how few certification fields the sample contains. I report the resulting p-value alongside, rather than in place of, the clustered standard error's p-value and I discuss any divergence between them directly in the results below rather than treating one as the default.

Interpreting the post-mandate change as the effect of the requirement rests on the assumption that, absent the mandate, certifications in SOR-required fields would have evolved in parallel to those in the comparison fields. While this assumption is inherently untestable, several pieces of evidence are consistent with it. First, Figure 3a shows that raw certification counts in the two groups track each other closely before the policy announcement. Second, I examine the event-study coefficients in the pre-announcement period and report a joint test that they are equal to zero. Even so, these diagnostic tests cannot rule out a smooth, ongoing pre-trend that continues into the post period. In Table A5, I assess the sensitivity of the pooled estimate to violations of parallel trends of varying magnitude following Rambachan and Roth (2023). The discussion below considers the most important remaining threats to the identifying assumption.

Two confounders deserve particular attention. The first is the COVID-19 pandemic, which arrived several months before the mandate took effect. If the pandemic differentially

[9]Wild cluster bootstrap procedures are a common alternative in linear settings with few clusters, but do not extend cleanly to the fixed-effects Poisson pseudo-maximum likelihood estimator used here, which lacks the well-defined residuals the standard wild bootstrap requires.

affected demand for or supply of teachers across certification fields—for example, by making elementary teaching relatively less attractive or shifting hiring toward non-SOR fields such as special education—the estimates could partly reflect those changes rather than the SOR requirement itself. The second is that the mandate coincided with a redesign of the content exams in the affected fields. When the standalone SOR exam was introduced, reading pedagogy content was removed from the EC–6 and ELAR 4–8 content exams, and revised versions of those exams were launched. Consequently, the estimates capture the combined effect of the new testing regime rather than the SOR requirement in isolation.

4.2 Screen: Does the Exam Identify Effective Teachers?

The welfare implications of the exam's supply effects depend on whether it screens for knowledge that matters for teaching effectiveness. If the exam identifies teachers who are more effective in the classroom, the costs documented above may be offset by gains in teacher quality among those who enter the profession. This section evaluates that possibility on three margins: whether exam scores predict teacher effectiveness overall, whether the passing threshold specifically separates more and less effective teachers, and whether a different threshold would meaningfully change the effectiveness of newly admitted teachers.

4.2.1 Predictive Validity of Certification Exam Scores

One way a certification exam may function as a screen is by providing information about teacher effectiveness before hire. A substantial literature has searched for such signals and largely come up short, though certification exam scores are among the few characteristics that consistently—if modestly—correlate with future effectiveness (Clotfelter et al., 2007; T. J. Kane et al., 2008; Rockoff et al., 2011). This section asks whether the SOR exam is such a signal, and whether it is a more informative one than a generic pedagogy exam like the PPR.

To assess whether certification exams predict teacher effectiveness, I estimate student-level value-added models of the form

$$A_{ijt} = \gamma \cdot \widetilde{S}_j + X'_{ijt}\beta + \bar{X}'_{cjt}\delta + \bar{X}'_{sjt}\phi + \mu_t + \lambda_g + \varepsilon_{ijt}, \tag{3}$$

where A_{ijt} is the standardized STAAR score of student i taught by teacher j in year t, and $\widetilde{S}_j$ is teacher j's first scaled score on the relevant certification exam, standardized within test takers. Controls include student prior achievement in reading and math (with quadratic and cubic terms), indicators for sex, race/ethnicity, limited English proficiency (LEP), special education, gifted, and economic disadvantage status, as well as classroom-level averages of these covariates. All specifications include year and grade fixed effects, with additional models including district and campus fixed effects. I estimate separate models for the SOR and Pedagogy and Professional Responsibilities (PPR) exams, and interpret γ as a signal of teaching quality observable before hire.

To test directly whether the SOR exam carries information beyond the PPR exam, I also estimate horse-race specifications that include both scores in the same model. A positive and statistically meaningful SOR coefficient after controlling for PPR would indicate the SOR exam captures information not already contained in the general pedagogy exam.

4.2.2 Regression Discontinuity in Value-Added at the Passing Threshold

While the predictive validity of the exam is informative about its overall relationship with effectiveness, it says nothing about whether the specific pass/fail line the policy draws falls where that relationship separates stronger from weaker candidates. To assess this, I examine whether the SOR passing threshold identifies differences in teacher effectiveness using a regression discontinuity design. Specifically, I estimate Equation 1 with the outcome replaced by teacher value-added.[10] In effect, this analysis asks whether the threshold used by policymakers meaningfully separates teachers who differ in subsequent effectiveness.

This design rests on comparing observed value-added across the threshold, but whether candidates make it to the classroom may depend on whether they pass the exam. If candidates who fall just below the passing margin and persist into the profession differ in unobservable ways from those who fail, the result of this RD could be biased. To account for this selection issue, I use the trimming procedure adapted from Lee (2009). I first estimate

[10] A teacher's value-added is estimated separately via equation 3 with $\widetilde{S}_j$ omitted and a teacher-level random effect included in its place. A teacher's value-added is thus the best linear unbiased predictor (BLUP) of this random effect. This procedure is similar to the commonly used empirical Bayes estimator in that it shrinks toward the sample mean in proportion to the precision of each teacher's estimate (i.e., teachers linked to fewer students are shrunk more towards the mean).

a regression discontinuity of whether I observe a teacher's value-added on exam score at the passing threshold. I use this as an estimate of the share of observed passers who would not have been observed had they instead failed.[11] I then rank observed passers by value-added and re-estimate the value-added discontinuity twice, once under each extreme assumption about who the marginal passers are. The upper bound drops the bottom share of the passing distribution equal to this excess (i.e., assuming the marginal passers are the weakest teachers). This is the largest value the effect of passing can take under any assumption about where the marginal passers fall in the value-added distribution. The lower bound drops the top share instead (i.e., assuming the marginal passers are the strongest teachers), giving the smallest value the effect can take. Reporting both bounds together shows whether the RD's conclusion depends on this selection concern.

4.2.3 Value-Added Under Counterfactual Passing Thresholds

The analysis above asks whether the threshold separates teachers of different levels of effectiveness. A complementary question is what a lower threshold would imply for the value-added of the teacher it would newly admit. To answer this, I estimate the relationship between value-added and score directly:

$$VA_j = \alpha + \beta \cdot S_j + \varepsilon_j, \tag{4}$$

where VA_j is teacher j's estimated value-added in reading (the same random-effects estimate used in the discontinuity analysis above) and S_j is the teacher's raw first-attempt SOR score. I estimate Equation 4 via OLS with robust standard errors and use the fitted line to predict the value-added of the marginal teacher at the actual threshold (240) and at three counterfactual thresholds (220, 200, and 180). Since the predictions come from the same linear fit, the difference between them is proportional to β, so a test of whether this difference is distinguishable from zero is equivalent to a test on the slope itself.

This exercise will be flawed if we incorrectly estimate the true conditional mean of teacher

[11] For example, if 25 percent of passers near the threshold are observed with a value-added score, compared to 22.5 percent of failers, the excess share is $(0.250 - 0.225)/0.250 = 0.10$, so roughly 10 percent of observed passers are treated as marginal for the purposes of the bounds.

value-added at points below the current cutoff. Three concerns bear on whether that is the case. The first is functional form: Equation 4 imposes a linear relationship between score and value-added over a wider range of scores than the exam was designed to distinguish, and the true relationship could curve or flatten below the threshold. I do not attempt to correct for this and instead treat the linear specification as a transparent starting point. The second and third concerns are both about who is included in the sample used to estimate this relationship. Since passing raises the probability a candidate is later observed teaching at all, the observed sample overrepresents higher performers among passers relative to failers (this is the same problem addressed via Lee bounds in the discontinuity analysis above). Relatedly, extending the estimated relationship to scores well below the current threshold requires either extrapolating from passers alone, who provide no data in that range, or including failers who nonetheless persisted into teaching, a group plausibly selected on traits (e.g., persistence or limited outside options), that may themselves correlate with effectiveness. I address both by estimating Equation 4 on two samples, one restricted to passers and one including all first-attempt test takers. I report the results from both and interpret the range as a plausible bound on the effectiveness cost of a lower threshold.

4.3 The Effect of Texas Reading Academies

The final analysis tests whether structured coursework can help candidates clear the bar without lowering it. To do this, I compare exam outcomes at EPPs that did and did not offer TRAs, before and after implementation, using a difference-in-differences design. Outcomes are the proportion of first-attempt test takers at each EPP who pass the SOR exam and the average first-attempt score, standardized within year.[12]

The event study specification is

$$Y_{it} = \alpha_i + \delta_t + \sum_{k \neq 2023} \beta_k \cdot \text{Academy}_i \times \mathbb{1}(t = k) + \varepsilon_{it}, \tag{5}$$

with EPP fixed effects α_i, year fixed effects δ_t, and standard errors clustered at the EPP

[12]Because additional EPPs adopted academies in 2025–26, and to avoid comparing already-treated to not-yet-treated programs under staggered adoption (Goodman-Bacon, 2021), I separately analyze the 2024–25 and 2025–26 cohorts against a never treated comparison group of EPPs.

level. The β_k measure academy–non-academy differences in year k relative to the year before implementation. A companion difference-in-differences model replaces the interaction terms with a single $\text{Academy}_i \times \text{Post}_t$ indicator.

Identification again rests on parallel trends. In this case, this means that absent the academies, outcomes at adopting and non-adopting EPPs would have evolved similarly. Two features of the design support the plausibility of this assumption for the 2024–25 cohort. First, the pre-period event study coefficients provide a test of differential trends before EPPs select into the academies. Second, adopting and non-adopting EPPs are balanced on pre-period pass rates and scores for both the SOR and PPR exams (Table 4). I focus the main analysis on this cohort for two further reasons. The 2025–26 cohort's pre-period event study coefficients reject the joint test of parallel trends, so a difference-in-differences comparison for this cohort is not well identified, and its post-adoption window covers only a partial school year, providing limited outcome data even setting the pre-trend concern aside. I report the 2025–26 cohort's results separately in Appendix Table A9 for completeness rather than as a corroborating estimate. Since I exclude one of the two academy cohorts from the main analysis, I interpret these results as preliminary rather than settled evidence on whether structured preparation affects exam performance.

5 Results

The SOR exam first-attempt pass rate hovers around 80 percent over the sample period, which is roughly comparable to the PPR exam and well above the EC-6 Content exam (Figure A1). The level is high enough that most candidates eventually pass. It is also low enough that one in five fails on the first attempt and faces exam fees, preparation time, and delayed certification. The analyses below test the extent to which the exam is a barrier at the individual- (Section 5.1) and market-level (Section 5.2), a screen (Section 5.3), and the extent to which preparation can mitigate any supply constraints (Section 5.4).

5.1 The Individual Cost of Failing

The most immediate consequence of failing is retaking. Table 5 reports that passing reduces the probability of retaking the exam within a year by 77.9 percentage points. The majority of candidates who narrowly fail make at least one additional attempt. The more substantive question is whether failure delays or prevents certification and employment. Passing increases the probability of earning a Standard teaching certificate within six months by 9.2 percentage points. By twelve months the estimate attenuates to 4.9 percentage points and loses statistical significance, which indicates that most candidates who narrowly fail eventually retake, pass, and certify. The friction is not fully absorbed by the certification process, however. Passing increases the probability of working as a teacher the following October by 6.5 percentage points. Notably, this effect is not mechanical. Individuals in Texas can and do teach without Standard certification, and roughly a third of test takers are already in classrooms before their exam (Table 1). Failing thus reduces employment through behavior rather than strict regulation, whether in district hiring decisions or in candidates' own willingness to persist. Figure 1 shows these discontinuities graphically.

These results are robust. Point estimates are nearly identical with and without month-of-exam fixed effects (Table A3) and stable across alternative data-driven bandwidth selectors (Table A4). Some estimates lose statistical significance without controls or under narrower bandwidths, but in each case the point estimates barely move reflecting that these decisions result in differences in precision, not in the qualitative result.

The overall results conceal a sharp divide by certification route (Table 6). Among traditional candidates, who are typically finishing university-based programs and not yet employed as teachers, the exam comes at the end of the pipeline. Those who narrowly fail retake at very high rates (92.4 percentage points), and passing at the margin raises six-month certification by 16.0 percentage points. These results are consistent with candidates sitting for the exam on the cusp of completing all other requirements. By twelve months the gap in the share of certified candidates above and below the cutoff closes entirely, with a small, wrong-signed, insignificant point estimate (-1.7 percentage points), and there is no effect on employment in the subsequent October. For a traditional candidate who just fails,

the exam delays a credential they were about to receive without derailing the career behind it.

The results for alternative route candidates tell another story. Even among alternative candidates who pass, fewer than a quarter certify within a year (Figure 2, panel (c)), far below every other route. The certification effect of passing is small at six months (2.5 percentage points) but grows to 10.6 percentage points at one year ($p < 0.001$), and narrowly failing reduces the probability of teaching the following October by 12.0 percentage points. This pattern is consistent with where the exam sits in the alternative route's sequence. Many alternative programs require candidates to pass the relevant TExES content and content-pedagogy exams before they can be placed on an intern certificate, so passing the exam is typically one of the first steps toward the classroom rather than one of the last, and failing it blocks entry to that pathway altogether.

The out-of-state and certification-by-exam routes fit the same interpretation. Failing produces a real certification penalty at six months (21.8 and 15.6 percentage points, respectively) that shrinks by one year, with no meaningful employment effect, though both samples are small enough that the estimates are imprecise. This pattern fits candidates who typically hold a teaching certificate already and are sitting for the SOR exam to add a field rather than to enter the profession. For these candidates, the exam sits after most of the work of becoming a teacher is already done.

These results show that the cost of narrowly failing depends on where the exam sits in a candidate's route to certification: at the end of a sequence for traditional candidates, at the entry point for alternative candidates, and after certification is already secured for out-of-state and by-exam candidates. But estimates at the passing margin capture only candidates who show up to test and just miss. If the exam also deters people from attempting at all, the fuller effect on supply appears not at the cutoff but in the flow of new certifications, the subject of the next section.

5.2 The Market-Level Cost

Panel (a) of Figure 3 shows the market-level pattern of newly issued standalone certifications in SOR-required and non-SOR fields by half-year. Before the mandate, the two series

move together. Both show a pronounced seasonal sawtooth, peaking each January–May as candidates completing traditional programs certify ahead of spring graduation, and drift gradually downward from 2016 through 2019, consistent with the broader national decline in entry into teaching over this period (Bartanen & Kwok, 2023; Kraft & Lyon, 2024). The COVID-19 pandemic mutes both series' January–May 2020 peak but disrupts neither series relative to the other. Against this backdrop, the divergence around the mandate is clear. Certifications in fields that would require the SOR exam surge in the second half of 2020, breaking the usual seasonal trough, as candidates rush to certify before the requirement took effect. The series then collapses once the exam became required in January 2021 and settles onto a persistently lower path. The gap between the two series widens at the mandate and has not closed through 2025.

Panel (b) formalizes this pattern by plotting the coefficients estimated from Equation 2. Each point represents the estimated rate ratio of certifications in SOR-required relative to non-SOR fields, normalized to the first half of 2019. Pre-announcement rate ratios hover around, and are statistically indistinguishable from, one, and a joint test fails to reject the null that the pre-period coefficients are jointly zero ($p = 0.649$; Table 7). The rate ratio rises to roughly 1.3 in the second half of 2020 as candidates accelerated certification ahead of the mandate, then falls to roughly 0.65 in the first half of 2021 as that pipeline exhausted. From 2022 onward, the coefficients range from 0.851 to 0.581. In other words, an EPP that would have issued one hundred certifications in SOR-required fields, given its own history and the statewide trend in non-SOR fields, now issues eighty-five to fifty-eight.

The difference-in-differences estimate in Table 7, Panel B, condenses these dynamics into a single number. Including the anticipatory period, certifications in SOR-required fields fell by 21.5 percent ($0.785 - 1 = -0.215$) relative to non-SOR fields within EPPs, significant at the 5% level under the clustered standard error and at the 10% level under randomization inference (RI $p = 0.081$). The donut specification, which excludes half-years 2019H2–2021H2 to remove the anticipatory surge and its immediate aftermath, implies a larger decline of 31.4 percent ($0.686 - 1 = -0.314$), significant at the 5% level under both the clustered standard error and randomization inference (RI $p = 0.045$). The two inference methods agree closely across most specifications, but diverge for the pooled and alternative-route

estimates specifically, both of which combine certification fields of markedly different size; this is consistent with cluster-robust asymptotics performing less reliably where cluster sizes are most heterogeneous, which is exactly the setting randomization inference is designed to address.[13]

Columns 2–5 of the same table show these results separately by certification route. The contraction spans every non-traditional route while mostly sparing the traditional one. Traditional programs show a small, statistically insignificant rate ratio of 0.926, consistent with the RD finding that a narrow failure is only a temporary setback for candidates embedded in university-based preparation. Certifications fell by roughly 22 percent in the alternative route, 31 percent out of state, and by nearly half in the certification-by-exam route.

These route-specific estimates warrant caution. The joint pre-period test rejects for three of the four routes (traditional, alternative, and by-exam), but the underlying pre-trend problems differ by route. For traditional and out-of-state candidates, the pre-period coefficients are tightly estimated but show no clear direction, hovering near one without a discernible trend, so the rejection for traditional reflects statistical precision rather than an economically meaningful pre-trend.[14] For alternative and by-exam candidates, the pre-period coefficients decline steadily in the run-up to the mandate, a genuine drift rather than noise, running in the same direction as the post-mandate effect. This pattern is most severe for by-exam, whose pre-period rate ratio falls from 1.306 to 1.085 over the six pre-periods. Randomization inference in the post-implementation window reinforces this split only partly. It is unfavorable for traditional (RI $p = 0.074$) as the pre-trend evidence would suggest, but favorable for alternative ($p = 0.002$) and by-exam ($p < 0.001$) despite their visible pre-trends, and also favorable for out-of-state ($p = 0.015$). Because a real pre-trend is a threat to identification that a favorable randomization test does not resolve, I do not treat the alternative and by-exam estimates as cleanly identified, and I read all four route-level results

[13] Sensitivity to violations of parallel trends is likewise strongest for the post-implementation period. Table A5 shows that, allowing the pre-period trend to continue linearly into the post period, the estimated decline over 2019H2–2025H2 is significant at the 10% but not the 5% level, while the estimated decline restricted to the post-implementation period, 2022H1–2025H2, is significant at the 5% level. Both estimates are sensitive to allowing the pre-period trend to bend beyond a small threshold. The weaker result for the full post-announcement period is consistent with its inclusion of the second-half-2020 anticipatory surge, whose sign is opposite that of the post-implementation decline.

[14] The joint test does not reject for out-of-state, but its pre-period coefficients are similarly flat.

as a rough decomposition of where the aggregate decline is concentrated rather than as four separately identified effects. The pooled estimate, whose own pre-trend is flat and whose joint test does not reject, remains the primary market-level finding.

One possible explanation for the decline is substitution, with candidates redirecting toward non-SOR fields rather than exiting or delaying. I find no evidence for this through two complementary analyses. First, at the passing margin, the regression discontinuity in Table A7 splits the certification outcome by field. Failing the exam sharply reduces the probability of certifying in an SOR-required field, by 10.6 percentage points within six months and 8.5 within a year, while the effect on certifying in a non-SOR field is small, wrong-signed, and statistically indistinguishable from zero. Candidates who narrowly fail do not pivot into unaffected fields. Second, the survival analysis in Figure A3 extends this logic beyond the margin to the full population of exam takers. Following individual candidates from their first attempt on *any* certification exam, post-mandate cohorts certify more slowly than pre-mandate cohorts overall and in SOR-required fields specifically, while their progress toward non-SOR certificates is essentially unchanged. Had candidates been substituting across fields, non-SOR certification would have risen to offset the SOR-required decline. It did not. The mandate's effect is a genuine slowdown in the affected fields, not a reshuffling across them.

5.3 Does the Exam Screen?

The supply costs documented above are the barrier half of the ledger. The screening half asks whether the exam measures knowledge that matters for student learning. Table 8 relates student learning gains on STAAR reading (Panel A) and math (Panel B) achievement exams to their teachers' SOR and PPR certification exam scores.

The reading results provide the most direct test of domain-specific screening. A one standard deviation increase in a teacher's SOR score is associated with a 0.011 standard deviation increase in student reading achievement in the baseline specification (column 1). The coefficient grows to 0.013 when comparisons are restricted to teachers within the same district (column 2) and school (column 3), which indicates the relationship is not driven by higher-scoring teachers sorting into higher-performing schools. All three estimates are

statistically significant. The corresponding PPR estimates are substantially smaller. They are 0.003 and insignificant at baseline and reach only 0.005 in the fixed-effects specifications (columns 4–6). SOR scores also predict math achievement, with estimates of 0.014 to 0.017 standard deviations across specifications (Panel B), somewhat larger than the reading estimates. This is consistent with the foundational role of literacy in academic learning generally, and with the broader finding that teachers and schools exert more influence over mathematics achievement than reading achievement, since much of children's literacy development occurs outside the classroom.[15] These magnitudes are modest in absolute terms but comparable to the broader certification literature (Clotfelter et al., 2007; Cowan et al., 2023).

The clearest evidence that the two exams measure different things comes from entering them together. Columns (7)–(9) include SOR and PPR scores in the same model, so each coefficient reflects the association net of the other. The SOR coefficient is essentially unchanged, at 0.009 to 0.012 standard deviations and significant across specifications, while the PPR coefficient is small, wrong-signed, and insignificant throughout (−0.006 to −0.001). Whatever predictive content the PPR score carried on its own is subsumed by the SOR score and the reverse is not true. The SOR exam captures instructionally relevant knowledge about reading that the general pedagogy exam does not, rather than proxying for a common component of teacher quality that both exams pick up.

While exam scores predict teacher effectiveness, the passing threshold does not appear to separate teachers of meaningfully different effectiveness in reading. Table 9 tests directly for a discontinuity in observed teacher value-added at the passing cutoff. The reading estimate is −0.004 (SE 0.019) and the math estimate is 0.029 (SE 0.031), both small and statistically indistinguishable from zero. Using the bounding procedure described in Section 4.2.2, I test whether this null survives accounting for the resulting selection into the value-added sample. In reading, the bounds are tight and remain centered near zero (−0.013 to 0.033), so the null does not appear to be an artifact of selection. In math, the bounds widen considerably (−0.022 to 0.090). Once selection into the value-added sample is taken into account, the

[15]This pattern is directly observable in these data. The standard deviation of teacher value-added is 0.20 student standard deviations in mathematics compared with 0.12 in reading, so there is less variation in teacher effectiveness for a reading-focused exam to detect.

math bounds admit a wide range of possibilities. Under the assumption that marginal passers are the weakest teachers, the threshold appears to admit meaningfully stronger teachers than it turns away, but under the assumption that marginal passers are the strongest, this gap disappears entirely.

That the threshold does not sort at its current location leaves open what would happen if it were set elsewhere. Figure 5 extends the estimated relationship between score and reading value-added below the cutoff to predict the value-added of the marginal teacher a state would admit under a lower threshold. Using the full sample of first-attempt test takers, moving the threshold from 240 to 220 implies a marginal teacher 0.009 standard deviations less effective than the one currently at the margin, growing to 0.026 standard deviations at a threshold of 180; each of these differences is statistically significant. Restricting the sample to passers, which avoids relying on failers who persisted into teaching but extrapolates into a range where no data are observed, produces similarly signed but smaller and statistically insignificant estimates (-0.005 to -0.014). The two specifications bracket the plausible cost of a lower threshold rather than pinning it down, but both point in the same direction. Lowering the bar would admit teachers who are, on average, somewhat less effective at raising reading achievement, even though the candidates immediately below the current cutoff are not distinguishable from those immediately above it.

These results suggest that the candidates who bear the supply costs are no less effective in reading than the ones who clear the bar just above them, consistent with evidence from other licensure settings (Orellana & Winters, 2025). Whatever value the exam has as a screen therefore appears to come from the additional predictive information the score carries, documented above, or by removing candidates further below the threshold insofar as they would make less effective teachers.

5.4 Can Preparation Lower the Barrier?

The results so far establish a tension. The exam is a meaningful barrier to teacher supply, but also provides a real signal of teacher effectiveness. One response to these results may be to lower the threshold. An alternative to this is to find ways to help more candidates clear it. In this section, I test whether Texas Reading Academies implemented at EPPs can do

this.

I focus on the ten EPPs that adopted academies in 2024–25, the first year they were offered, comparing their first-attempt SOR outcomes to those of never-adopting programs before and after adoption. A second wave adopted in 2025–26, but I exclude them from the main analysis for two reasons. First, they fail parallel trends, entering on a declining performance path relative to non-adopters in the years before adoption (Table A9). Second, their post-adoption window covers only part of a school year (data are only available through March 2026). These programs lack a quality comparison group and have an insufficient amount of outcome data.

For the first cohort, by contrast, the pre-period trends are clean (Figure 6). Academy and non-academy programs moved together on both SOR pass rates and scores before adoption, and the pre-period event-study coefficients are individually and jointly insignificant (Table 10; joint $p = 0.529$ for pass rates, $p = 0.544$ for scores). The two groups then diverge in 2024–25. Academy programs post a 4.0 percentage point gain in first-attempt SOR pass rates (column 1) and a 0.15 standard deviation gain in scores (column 2), with a difference-in-differences pass-rate estimate of similar size (columns 5 and 6). The gap holds into the partial 2025–26 year, at 5.4 percentage points and 0.18 standard deviations, so the advantage does not fade as the first candidates move through.

To check whether the gain is specific to SOR content, I turn to the PPR exam, which includes content the academies do not cover. Literacy-specific preparation should raise SOR performance more than PPR, and the results are mixed. The PPR pass rate rises 4.2 percentage points in the adoption year (column 3), close to the SOR gain, but falls to a null in the partial 2025–26 year, and the pooled difference-in-differences estimates are smaller than the SOR gains and insignificant (columns 7 and 8). The PPR movement is thus confined to the first year rather than sustained, which cuts against the reading that adopting programs were simply improving across the board. Some spillover is plausible in any case, since coursework on explicit, systematic instruction may sharpen general pedagogical skill.

Taken together, the results suggest the academies help exam performance, but they fall short of proving it. Candidates at EPPs in the first-cohort of academy adopters improved on

the SOR exam, but a similar first-year gain on the PPR exam means I cannot fully rule out that these programs were improving in more general ways. And the second cohort, which lacks clean pre-trends and a full year of data, cannot corroborate the first. What the design can establish is that adopting programs were on parallel trends beforehand and pulled ahead on the tested content after, which is consistent with the academies working.

6 Discussion

Occupational licensing may operate as a *screen* on quality but can also act as a *barrier* and restrict supply (M. M. Kleiner, 2000). Applied to teacher certification exams, prior work has largely found that teacher certification exams raise significant barriers and provide little value as screens. Angrist and Guryan (2008) showed that states which introduced stricter certification requirements reduced teacher supply and raised wages, but did not alter teacher quality. More recent evidence from Deneault et al. (2025) found that increasing the difficulty of a teacher certification exam did not shift the distribution of teacher value-added. Given this evidence, the natural prior for a new certification exam is that its cost as a barrier will outweigh its benefits as a screen. This paper's results complicate that prior, but fall short of overturning it. Texas' Science of Reading exam is a substantial barrier to teacher supply, but it also provides valuable information about teacher effectiveness before entry to the classroom, and reducing the requirement would likely admit individuals into the profession who are, on average, less effective at raising student achievement in reading.

The SOR exam acts as a barrier across two margins. Narrowly failing the exam delays certification and reduces the probability of teaching the following year. Although this is somewhat mechanical, it is noteworthy in a context where 15.6% of teachers enter the classroom completely uncertified (Ghazzawi et al., 2025). This finding is heterogeneous by certification route, a dimension prior literature has not documented. For candidates close to certification, such as those coming from university-based programs or who have already earned a certification in another area or state, narrow failure is merely a delay. For alternatively certified candidates, the same failure cuts the probability of teaching the following October by twelve percentage points. This likely occurs because passing is what unlocks the

internship that lets them enter the classroom, and often the alternative route altogether. This is particularly relevant as an increasingly large share of new teachers come from alternative routes (National Council on Teacher Quality, 2025).

The market-level results extend the barrier to candidates further from the cutoff and those who avoid the exam altogether. Prior work on teacher licensure has documented the consequences of raising a passing standard or introducing testing where none existed, but not the effect of layering an additional exam onto an existing certification pipeline. This result is particularly relevant to states adopting Science of Reading examination requirements. New certifications in fields that came to require the exam fell substantially relative to comparison fields within the same programs and have not recovered five years later. This decline survives allowing for pre-existing certification trends to continue into the post-period. In addition, two checks rule out substitution: candidates who narrowly fail do not pivot into unaffected fields, and post-mandate cohorts do not certify faster in unaffected fields to compensate. The requirement slowed the flow of teachers into these fields rather than redirecting it. This cost reduction in teacher supply is particularly salient at a moment when interest in teaching is already falling (Bartanen & Kwok, 2023; Kraft & Lyon, 2024).

Despite the considerable supply costs of the Science of Reading exam, this study pushes back against a growing consensus for loosening or eliminating teacher certification requirements altogether (Hanushek & Rivkin, 2007). SOR scores predict student reading achievement two to four times more strongly than scores on the general pedagogy exam, and this result holds when both exams are entered in the same regression. This result is notable given two decades of research that have struggled to consistently identify pre-hire measures that correlate with teacher value-added (Clotfelter et al., 2007; Cowan et al., 2023; Goldhaber, 2007; Goldhaber et al., 2017; T. J. Kane et al., 2008; Rockoff et al., 2011). The results in this study show that an exam tied directly to instructional practices with strong evidence of raising student reading achievement is a considerably better predictor of teacher quality than one testing general pedagogical knowledge.

Still, I find little evidence that the exam discriminates between more or less effective candidates at the pass/fail threshold. Teachers who barely passed and barely failed have statistically indistinguishable reading value-added, and this null survives correcting for the

selection induced by passing itself raising the odds a candidate is later observed teaching. There is little theoretical reason to have expected otherwise. If score is a noisy proxy for continuously distributed latent teaching ability, nothing about crossing an arbitrary point on the score distribution should produce a jump. As corroborating evidence, Orellana and Winters (2025) build their design around this logic in the context of Connecticut's Praxis II and find that failing disproportionately deters candidates who would have been at least as effective as those who passed.

However, lowering the bar for certification is probably not without cost. Extending the estimated relationship between score and reading value-added below the current threshold implies a statistically significant effectiveness cost to admitting teachers at a lower cutoff. This is consistent with Larsen et al. (2026), who compare states that tightened their teacher licensing requirements to those that did not and find that tightening mostly removes candidates from the bottom of the quality distribution. If raising a requirement screens out weaker candidates, lowering one should admit them.

Given the current climate of education in the United States, it is very likely that policymakers will want to maintain minimum competency standards using certification exams. Roughly 40 percent of fourth graders now score below basic on the National Assessment of Educational Progress, the highest share since 2002 (National Center for Education Statistics, 2025). Teacher quality is among the most consequential school-level determinants of student achievement (Chetty et al., 2014; Rivkin et al., 2005), and recent evidence from relaxed pandemic-era entry pathways suggests that loosening requirements lowered it (Backes et al., 2024). Moreover, certification requirements are among the few levers states directly control in responding to this trajectory and, as noted, 20 states are now using them for exactly that purpose (National Council on Teacher Quality, 2024).

Whatever the case for a minimum standard, the current pass/fail margin is a costly way to enforce it. The supply cost is large, persistent, and concentrated on the routes the profession increasingly depends on, while the screening benefit sits in the score rather than at the cutoff. Texas may also have been fortunate in where this cost landed. Chronic shortages concentrate in mathematics, science, and special education (Cowan et al., 2016), none of which the requirement touches. A state imposing the same barrier in a genuine shortage

field would have far less slack to absorb it, and would do well to expand supply into those fields before restricting entry.

Both of these observations point toward the same underemphasized alternative, which is doing more with the exam than using it as a pass/fail mechanism. On the preparation side, the literature on certification testing has focused overwhelmingly on consequences for entry, and comparatively little on what programs and states can do to help candidates meet the standard. That gap, along with evidence that educator preparation programs are largely indistinguishable in their ability to produce effective teachers (von Hippel & Bellows, 2018), suggests real room to improve how we prepare teachers. The early evidence on Texas Reading Academies is consistent with this possibility, though a similar first-year gain on the pedagogy exam and a second cohort that fails its own pre-trend test mean it should be read as suggestive rather than settled. On the information side, the exam produces one of the stronger pre-hire signals of teaching quality available, but not at the pass-fail margin. Preparation programs could use scores to target support toward struggling candidates, and states could use them in addressing the persistent sorting of stronger teachers away from disadvantaged schools (Lankford et al., 2002). While neither approach removes the barrier the exam creates, both treat the exam as a source of information about teacher quality rather than only a gate to a credential.

6.1 Limitations

Several limitations bound these conclusions. The regression discontinuity estimates are local to the passing margin. The market-level analysis attempts to say more, but in turn, rests on a small number of clusters, shows some evidence of a pre-trend, and cannot fully separate the SOR requirement from a concurrent redesign of the affected content exams or from pandemic-era disruptions to the broader teacher labor market. The value-added estimates establish predictive validity rather than a causal effect of teacher knowledge on student achievement, since higher-scoring teachers may differ from lower-scoring peers in unobserved ways. All analyses using teacher value-added are further limited by the fact that candidates who score below the cutoff are less likely to work as teachers and therefore less likely to be observed at all. The counterfactual threshold analysis additionally assumes the

relationship between score and value-added is linear across a range of scores where relatively few teachers are observed, and predictions further from the current cutoff should be treated as correspondingly less certain. Finally, this is one exam in one state, in fields without chronic shortages. The costs and screening value estimated here need not generalize to a different exam, a different subject, or context with significantly different labor market conditions.

7 Conclusion

The evidence indicates that Texas' Science of Reading exam reshaped the teacher pipeline in ways that matter for both supply and quality. The requirement imposed real costs on candidates, especially those entering through alternative routes, and contributed to a persistent decline in new certifications in affected fields. Yet the exam also identifies meaningful variation in teacher effectiveness. SOR scores predict reading value-added substantially more strongly than existing pedagogy exams, while the pass/fail threshold itself does not separate more effective teachers from less effective ones. These patterns suggest that the exam's value lies in the information embedded in the score rather than in the cutoff that currently governs entry. Early evidence from Texas Reading Academies shows that structured preparation can raise pass rates, pointing to a path for states that wish to maintain minimum competency standards without exacerbating shortages. As more states adopt Science of Reading requirements, the evidence from Texas underscores the need to pair new licensure hurdles with investments in preparation that help candidates meet them.

References

Amin, R. (2022, April). New york officials vote to scrap edtpa teacher certification exam [Accessed: 2026-01-16].

Anglin, K. (2024). The Role of State Education Regulation: Evidence From the Texas Districts of Innovation Statute. *Educational Evaluation and Policy Analysis*, *46*(3), 534–554.

Angrist, J. D., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, *27*(5), 483–503.

Backes, B., Cowan, J., Goldhaber, D., & Theobald, R. (2024). Four years of pandemic-era emergency licenses: Retention and effectiveness of emergency-licensed Massachusetts teachers over time. *Economics of Education Review*, *101*, 102562.

Bartanen, B., & Kwok, A. (2023). From Interest to Entry: The Teacher Pipeline From College Application to Initial Employment. *American Educational Research Journal*, *60*(5), 941–985.

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, First Edition* [Backup Publisher: National Bureau of Economic Research Type: Book]. NBER.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. *The Future of Children*, *17*(1), 45–68.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs: Robust Nonparametric Confidence Intervals. *Econometrica*, *82*(6), 2295–2326.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51.

Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation Testing Based on Density Discontinuity. *The Stata Journal*, *18*(1), 234–261.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679.

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

Cowan, J., Goldhaber, D., Hayes, K., & Theobald, R. (2016). Missing Elements in the Discussion of Teacher Shortages. *Educational Researcher*, *45*(8), 460–462.

Cowan, J., Goldhaber, D., Jin, Z., & Theobald, R. (2023). Assessing Licensure Test Performance and Predictive Validity for Different Teacher Subgroups [Publisher: SAGE PublicationsSage CA: Los Angeles, CA]. *American Educational Research Journal*.

Deneault, C., Riehl, E., & Zou, J. (2025). *The unequal impacts of teacher certification exams on prospective teachers and students*.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393–447.

Ellis, C., Holston, S., Drake, G., Putman, H., Swisher, A., & Peske, H. (2023). Teacher prep review: Strengthening elementary reading instruction [Accessed January 2026].

Ghazzawi, D., Olofson, M., Landa, J., & Eluru, M. (2025, March). *Uncertified teacher rates 2019–20 through 2024–25* (tech. rep.). Texas Education Agency.

Goldhaber, D. (2007). Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness? *The Journal of Human Resources*, *42*(4), 765–794.

Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, *68*(4), 377–393.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Hankins, D. K. (2026). How to Embrace the 'Science of Math' Without Abandoning Your Existing Curriculum. *Education Week*.

Hanushek, E. A., & Rivkin, S. G. (2007). Pay, Working Conditions, and Teacher Quality. *The Future of Children*, *17*(1), 69–86.

Kane, T., & Staiger, D. (2008, December). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation* (tech. rep. No. w14607). National Bureau of Economic Research. Cambridge, MA.

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

Kirksey, J. J. (2024, July). *Amid Rising Number of Uncertified Teachers, Previous Classroom Experience Proves Vital in Texas. Policy Brief. No. 1* (tech. rep.) (ERIC Number: ED659927). Center for Innovative Research in Change, Leadership, and Education.

Kirksey, J., & Gottlieb, J. (2023). Teacher preparation goes virtual in the wild west: The impact of fully-online teacher preparation in Texas.

Kleiner, M. (2006). Licensing Occupations: Ensuring Quality or Restricting Competition? *Upjohn Press*.

Kleiner, M. M. (2000). Occupational Licensing. *Journal of Economic Perspectives*, *14*(4), 189–202.

Kleiner, M. M., & Krueger, A. B. (2013). Analyzing the Extent and Influence of Occupational Licensing on the Labor Market. *Journal of Labor Economics*, *31*(S1), S173–S202.

Kraft, M. A., & Lyon, M. A. (2024). The Rise and Fall of the Teaching Profession: Prestige, Interest, Preparation, and Satisfaction over the Last Half Century.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62.

Larsen, B., Ju, Z., Kapor, A., & Yu, C. (2026). The effect of occupational licensing stringency on the teacher quality distribution [Conditionally accepted; NBER Working Paper No. 28158]. *American Economic Journal: Economic Policy*.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*(3), 1071–1102.

MacKinnon, J. G., & Webb, M. D. (2020). Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, *218*(2), 435–450.

McArthur, G., Sheehan, Y., Badcock, N. A., Francis, D. A., Wang, H.-C., Kohnen, S., Banales, E., Anandakumar, T., Marinus, E., & Castles, A. (2018). Phonics training for English-speaking poor readers. *Cochrane Database of Systematic Reviews*, (11), CD009115.

Moats, L. C. (2020). Teaching reading is rocket science: What expert teachers of reading should know and be able to do. In *Reading research anthology: The why? of reading instruction* (pp. 29–73). Arena Press.

National Center for Education Statistics. (2025). *2024 NAEP reading assessment: Results at grades 4 and 8 for the nation, states, and districts* (The Nation's Report Card No. NCES 2024218). U.S. Department of Education, Institute of Education Sciences.

National Council on Teacher Quality. (2024). *Licensure tests - state teacher policy database* [Accessed: 2026-01-16]. National Council on Teacher Quality.

National Council on Teacher Quality. (2025). *What's the alternative? how some alternative certification programs are shortchanging teachers and students* (tech. rep.) (Federal Title II data analyzed in the report show that enrollment in alternative certification programs has increased by 63% since 2010–11.). National Council on Teacher Quality.

NJ State Policy Lab. (2025, January). Nj ends requirement for praxis exams for teacher certification [Accessed: 2026-01-16. New Jersey eliminated the basic skills (Praxis Core) teacher certification test effective January 1, 2025, to address teacher shortages and reduce barriers to entry.].

Novicoff, S., & Dee, T. S. (2025, January). *The achievement effects of scaling early literacy reforms* (EdWorkingPaper No. 23-887). Annenberg Institute at Brown University.

Orellana, A., & Winters, M. A. (2025). Licensure Tests and Teacher Supply. *Journal of Human Resources*, 0324–13465R2.

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, *90*(5), 2555–2591.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458.

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

Santos Silva, J. M. C., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, *88*(4), 641–658.

Seidenberg, M. S. (2017). *Language at the speed of sight: How we read, why so many can't, and what can be done about it*. Basic Books.

Spear-Swerling, L., & Brucker, P. O. (2004). Preparing novice teachers to develop basic reading and spelling skills in children. *Annals of Dyslexia*, *54*(2), 332–364.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374.

Spencer, N. (2024). Comprehensive early literacy policy and the "Mississippi Miracle". *Economics of Education Review*, *103*, 102595.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, *24*(3), 97–118.

Stiglitz, J. (1975). The Theory of "Screening," Education, and the Distribution of Income. *American Economic Review*, *65*(3), 283–300.

Swisher, A. (2022, July). Setting sights lower: States back away from elementary teacher licensure tests.

Torgesen, J. K. (2004). Preventing early reading failure. *American Educator*, *28*(3), 6–9.

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*, *64*, 298–312.

Tables and Figures

Table 1: Summary Statistics for First-Attempt Science of Reading Exam Takers

	Full Sample			Passed			Failed		
	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*
Passed SOR Exam	0.805	0.396	21692	1.000	0.000	17459	0.000	0.000	4233
Scaled Score on SOR Exam	251.612	17.297	21692	258.116	10.272	17459	224.784	14.272	4233
Certification Route									
Traditional	0.343	0.475	21692	0.354	0.478	17459	0.294	0.456	4233
Alternative	0.395	0.489	21692	0.363	0.481	17459	0.527	0.499	4233
By Exam	0.111	0.314	21692	0.116	0.320	17459	0.089	0.285	4233
Out of State	0.152	0.359	21692	0.167	0.373	17459	0.089	0.285	4233
Teaching before SOR Exam	0.315	0.465	21643	0.316	0.465	17419	0.311	0.463	4224
Demographics									
Has demographic information	0.645	0.479	21692	0.633	0.482	17459	0.692	0.462	4233
Female	0.921	0.269	13986	0.930	0.256	11056	0.890	0.313	2930
Non-white	0.429	0.495	14503	0.402	0.490	11496	0.530	0.499	3007
Outcomes									
Retake Exam	0.142	0.349	21692	0.000	0.008	17459	0.726	0.446	4233
Earn Standard Cert. in Six Months	0.432	0.495	21692	0.487	0.500	17459	0.204	0.403	4233
Earn Standard Cert. in One Year	0.588	0.492	21692	0.657	0.475	17459	0.306	0.461	4233
Employed as a Teacher next October	0.708	0.455	21692	0.738	0.439	17459	0.581	0.499	4233

Notes: Sample includes all first-attempt takers of the TExES Science of Reading exam (exam 293) from January 1, 2021 through October 23, 2023, restricted to candidates with valid scaled scores. Passed and Failed columns split the sample by first-attempt pass status (score $\geq$ 240). Demographic information is available for candidates who can be linked to PEIMS K–12 student records or THECB postsecondary enrollment records; the reduced sample sizes for Female and Non-white reflect this linkage. Certification outcomes measure whether the candidate earns any standard teaching certificate within the specified window. Employment is measured as of the last Friday in October following the exam date.

Table 2: Descriptive Statistics for New Certifications

	Pre-Mandate		Post-Mandate	
	Mean	N	Mean	N
SOR-Required Fields	0.417	139,575	0.360	203,241
Non-SOR Fields	0.583	139,575	0.640	203,241
Route				
Traditional	0.240	139,575	0.227	203,241
Alternative	0.338	139,575	0.335	203,241
By Exam	0.323	139,575	0.302	203,241
Out of State	0.099	139,575	0.136	203,241
Demographics				
Female	0.747	139,575	0.757	203,241
White	0.592	139,575	0.552	203,241
Hispanic	0.251	139,575	0.276	203,241
Black	0.108	139,575	0.117	203,241

Notes: Table reports descriptive statistics for the certification-level sample underlying the market-level analysis, prior to collapsing to the educator preparation program (EPP) × program-type × certification-field × half-year panel. Sample includes all newly issued standalone teaching certifications in the SOR-required and comparison fields listed in Table A2, 2016 through 2025. Post-mandate is defined as certifications issued from 2019H2 onward.

Table 3: Structure of the Market-Level Analytic Panel

	SOR-Required	Comparison
Certification fields	3	16
EPP-program-type units	200	200
Panel cells (EPP × field × half-year)	9,585	51,120
Share of cells zero-filled	51.2%	69.9%
Total certifications in panel	131,349	211,467

Notes: This table describes the panel formed by collapsing certification records in Table 2 to counts by educator preparation program (EPP) × program-type × certification-field × half-year, spanning 2016H1 through 2025H2. Certification fields are the 19 individual certification categories classified as SOR-required or comparison fields listed in Table A2. EPP-program-type units are the distinct combinations of educator preparation program and program type (e.g., Traditional, Alternative) observed issuing certifications. Panel cells are EPP-program-type × certification-field × half-year observations; a cell is coded zero if the EPP-program-type issued no certifications in that field during that half-year but was otherwise active, defined as having issued a certification in any field during that half-year. Periods before an EPP-program-type's first issued certification, or after its last, are excluded from the panel rather than coded as zero.

Table 4: Pre-Period EPP Characteristics by SOR Academy Status

	Academy EPPs		Non-Academy EPPs		Difference	
	Mean	SD	Mean	SD	Mean	SE
Certification Exams						
N Exam Takers	298.600	266.021	469.918	1616.239	171.318	(513.905)
Pass SOR Exam	0.871	0.171	0.908	0.080	0.037	(0.030)
SOR Exam Scaled Score	252.855	10.376	252.827	6.960	-0.027	(2.442)
Pass PPR Exam	0.860	0.095	0.862	0.142	0.002	(0.046)
PPR Exam Scaled Score	259.028	8.275	258.847	8.303	-0.181	(2.757)
Program Type						
Traditional	0.733	0.411	0.558	0.472	-0.175	(0.155)
Alternative	0.267	0.411	0.412	0.467	0.145	(0.154)
By Exam	0.000	0.000	0.025	0.150	0.025	(0.048)
Out of State	0.000	0.000	0.005	0.051	0.005	(0.016)
Certs Issued						
Certs. Requiring SOR Exam	0.491	0.090	0.509	0.171	0.018	(0.055)
Certs. Not Requiring SOR Exam	0.509	0.090	0.491	0.171	-0.018	(0.055)
Demographics						
Female	0.810	0.075	0.784	0.098	-0.026	(0.032)
Asian	0.013	0.008	0.023	0.028	0.010	(0.009)
Black	0.047	0.040	0.121	0.193	0.073	(0.061)
Other	0.022	0.010	0.023	0.017	0.001	(0.005)
Hispanic	0.427	0.295	0.405	0.257	-0.023	(0.087)
White	0.490	0.267	0.429	0.248	-0.062	(0.083)
N EPPs	10		98		108	

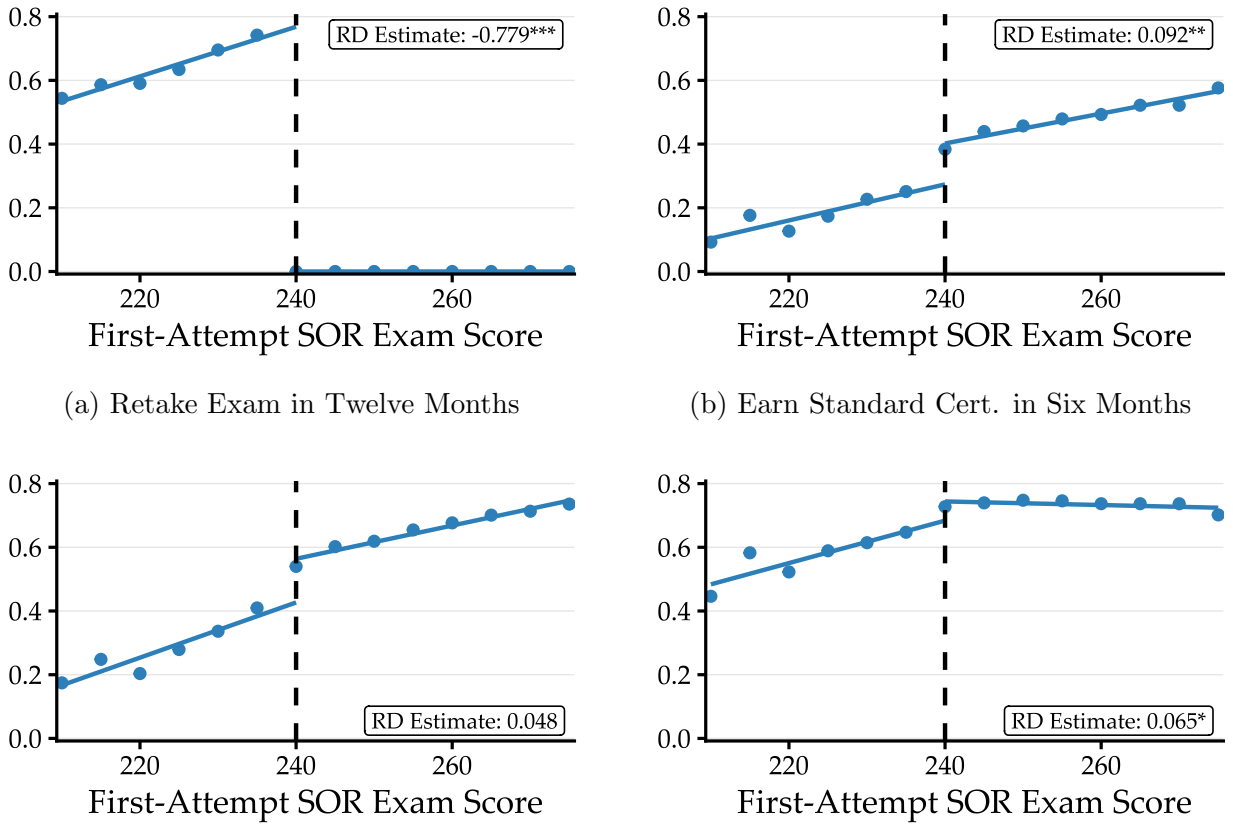
Notes: Table compares pre-period characteristics of Educator Preparation Programs (EPPs) authorized to offer SOR Academy coursework beginning in 2024–25 to EPPs that never adopted SOR Academies during the study period. Means are calculated over academic years 2021–22 through 2023–24 (pre-academy period). The difference column reports the academy minus non-academy difference with standard errors from a two-sample t-test in parentheses. EPPs that adopted SOR Academies in 2025–26 are excluded from both groups to avoid contamination from anticipatory changes at programs preparing to adopt. See Table A8 for the full list of EPPs and adoption years.

† Demographics presented for individuals who earn Standard teaching certificates associated with an EPP.
$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 5: RD Estimates of Passing the SOR Exam

	(1) Retake SOR Exam	(2) Earn Std. Cert. in 6 Months	(3) Earn Std. Cert. in 1 Year	(4) Teach Next October
Conventional	-0.777^{***}	0.0932^{***}	0.0592^{*}	0.0640^{*}
	(0.0191)	(0.0242)	(0.0266)	(0.0250)
Bias-corrected	-0.779^{***}	0.0923^{***}	0.0485	0.0650^{**}
	(0.0191)	(0.0242)	(0.0266)	(0.0250)
Robust	-0.779^{***}	0.0923^{**}	0.0485	0.0650^{*}
	(0.0230)	(0.0293)	(0.0301)	(0.0304)
Month of Exam FE	✓	✓	✓	✓
N	21,692	21,692	21,692	21,692

Notes: Table reports regression discontinuity estimates of the effect of passing the Science of Teaching Reading (SOR) exam (TExES 293) on teacher outcomes. Each column presents a separate outcome. Columns (1) shows the probability of retaking the SOR exam within one year of the exam date. Columns (2) and (3) show the probability of earning a standard teaching certificate within six months and one year of the exam date, respectively. Column (4) shows the probability of working as a teacher in Texas public schools on the last Friday of October following the exam date. All specifications include month-of-exam fixed effects. All estimates use a local linear regression with triangular kernel weights and data-driven bandwidth selection. Standard errors in parentheses. Conventional and bias-corrected rows report heteroskedasticity-robust standard errors; robust rows report the bias-corrected robust standard errors of Calonico et al. (2014), which account for bias in the local polynomial estimator near the boundary. Sample includes all first-attempt test takers of the SOR exam from January 2021 to October 2023.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$



(a) Retake Exam in Twelve Months

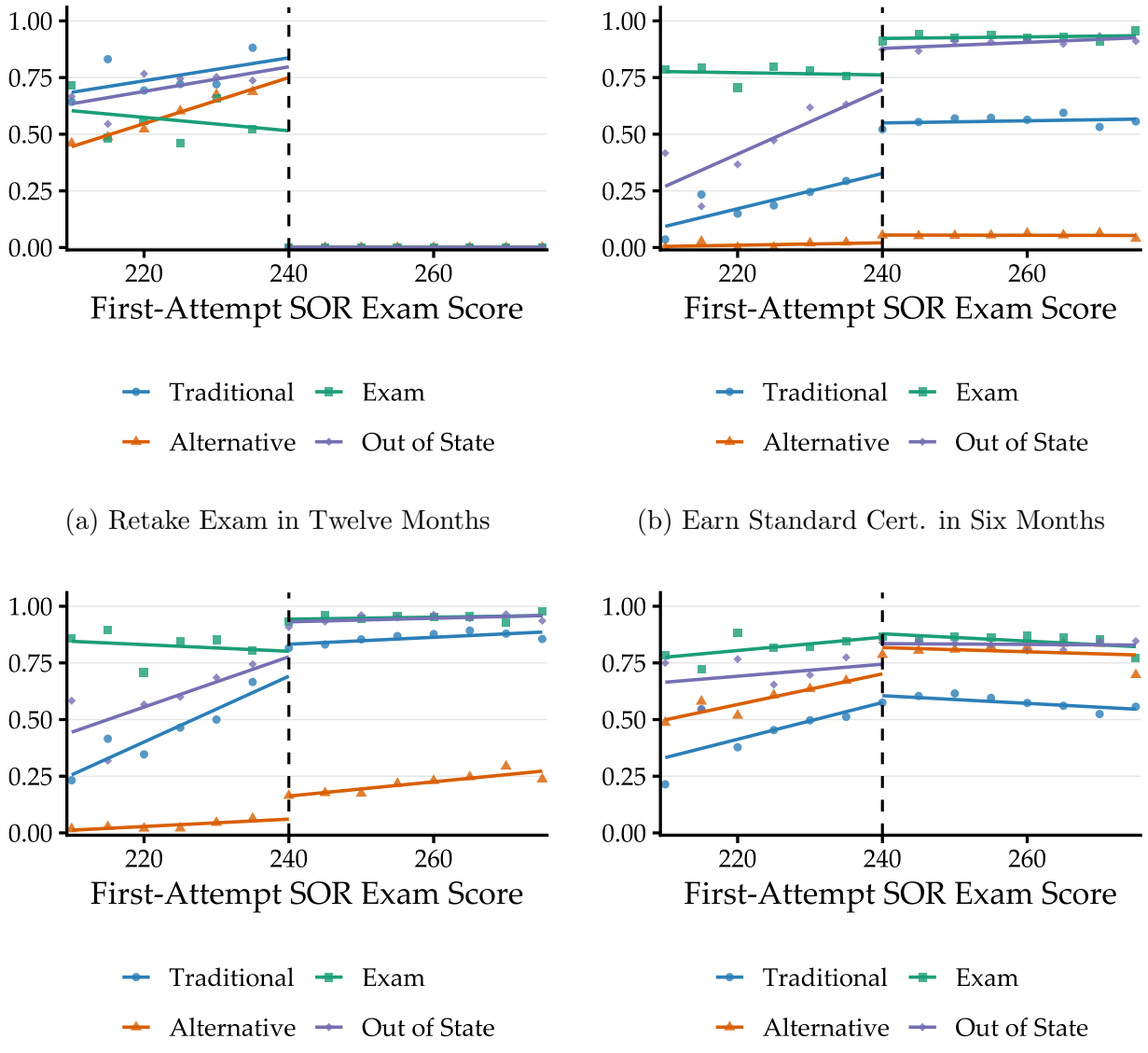
(b) Earn Standard Cert. in Six Months

(c) Earn Standard Cert. in Twelve Months

(d) Work as a Teacher Next October

Figure 1: Teacher Outcomes by Science of Reading Exam Score

Notes: Each panel plots outcomes for first-attempt test takers of the TExES Science of Reading (SOR) exam (exam 293) from January 1, 2021 to October 23, 2023. X-axis shows first-attempt exam scores grouped into five-point bins (e.g., scores 220–224 are assigned to bin 220); points represent the mean outcome proportion within each bin. Lines show separate linear fits on each side of the passing threshold of 240 (indicated by the vertical dashed line). Panel (a) shows the proportion who retake the SOR exam within one year. Panel (b) shows the proportion who earn a Standard teaching certificate within six months of the exam date. Panel (c) shows the proportion who earn a Standard teaching certificate within twelve months of the exam date. Panel (d) shows the proportion who work as a teacher in Texas public schools on the last Friday of October following the exam date. To protect individual privacy, cells with proportions below one percent are coded as zero and cells above 95 percent are coded as one. The annotated RD estimate in each panel is the bias-corrected robust regression discontinuity estimate of Calonico et al. (2014), using a local linear regression with triangular kernel weights and data-driven bandwidth selection, corresponding to the robust estimates reported in Table 5.
$*p < 0.05$, $**p < 0.01$, $***p < 0.001$



(a) Retake Exam in Twelve Months

(b) Earn Standard Cert. in Six Months

(c) Earn Standard Cert. in Twelve Months

(d) Work as a Teacher Next October

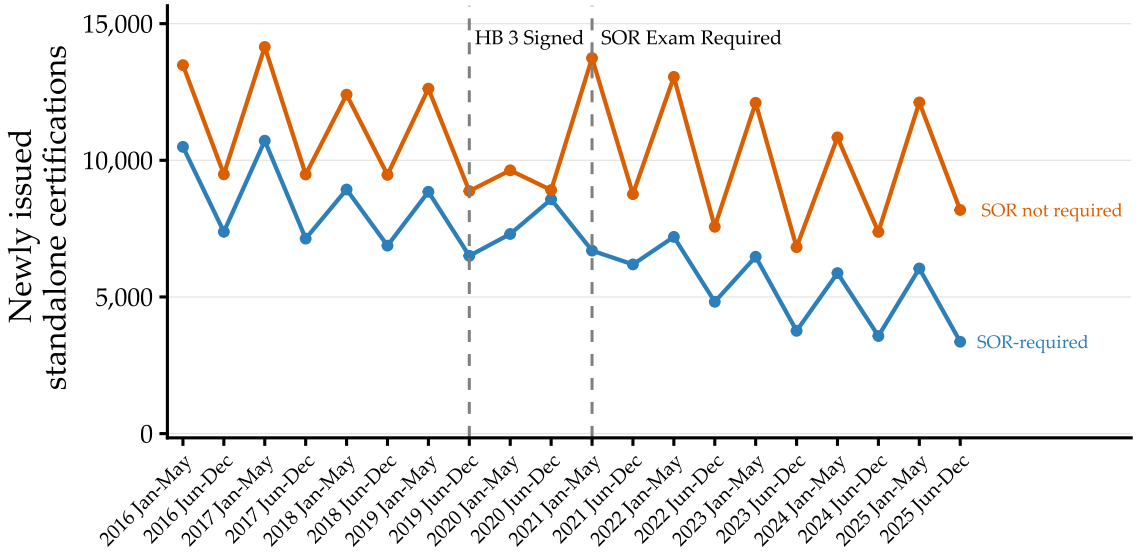
Figure 2: Teacher Outcomes by SOR Exam Score and Certification Route

Notes: Each panel plots outcomes by certification route for first-attempt test takers of the TExES Science of Reading (SOR) exam (exam 293) from January 1, 2021 to October 23, 2023. Certification routes are Traditional (university-based educator preparation programs), Alternative (alternative certification providers), By Exam (candidates certified by examination without a preparation program), and Out of State (candidates with out-of-state credentials). X-axis shows first-attempt exam scores grouped into five-point bins (e.g., scores 220–224 are assigned to bin 220); points represent the mean outcome proportion within each bin. Lines show separate linear fits on each side of the passing threshold of 240 (indicated by the vertical dashed line). Panel (a) shows the proportion who retake the SOR exam within one year. Panel (b) shows the proportion who earn a Standard teaching certificate within six months of the exam date. Panel (c) shows the proportion who earn a Standard teaching certificate within twelve months of the exam date. Panel (d) shows the proportion who work as a teacher in Texas public schools on the last Friday of October following the exam date. To protect individual privacy, cells with proportions below one percent are coded as zero and cells above 95 percent are coded as one.

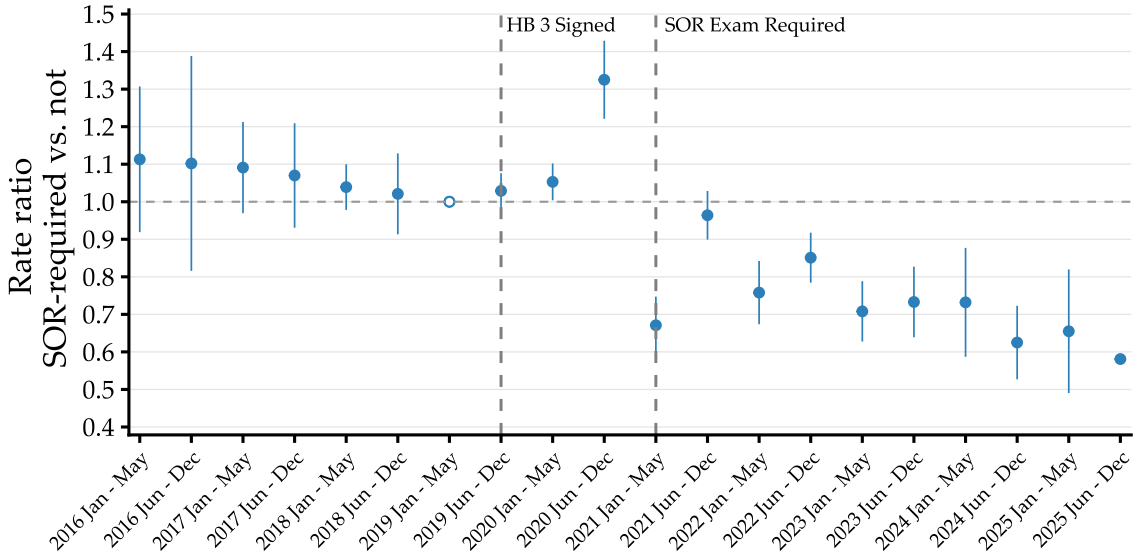
Table 6: RD Estimates of Passing the SOR Exam, by Certification Route

	(1) Retake SOR Exam	(2) Earn Std. Cert. in 6 Months	(3) Earn Std. Cert. in 1 Year	(4) Teach Next October
Traditional	-0.924^{***}	0.160^{***}	-0.017	0.0267
	(0.0398)	(0.048)	(0.058)	(0.0533)
Alternative	-0.681^{***}	0.025	0.106^{***}	0.120^{**}
	(0.0388)	(0.014)	(0.025)	(0.0443)
By Exam	-0.553^{***}	0.160	0.139	0.0281
	(0.0844)	(0.103)	(0.095)	(0.0908)
Out of State	-0.781^{***}	0.218^{**}	0.097	-0.0442
	(0.0598)	(0.073)	(0.068)	(0.0719)

Notes: Table reports regression discontinuity estimates of the effect of passing the Science of Teaching Reading (SOR) exam (TExES 293) on teacher outcomes, estimated separately for each certification route. Each cell reports the estimate from a separate regression. Column (1) shows the probability of retaking the SOR exam within one year. Columns (2) and (3) show the probability of earning a standard teaching certificate within six months and one year of the exam date, respectively. Column (4) shows the probability of working as a teacher in Texas public schools on the last Friday of October following the exam date. All specifications include month-of-exam fixed effects, which substantially improve precision. Estimates use local linear regression with triangular kernel weights and data-driven bandwidth selection. Estimates are bias-corrected and standard errors are the robust standard errors of Calonico et al. (2014), reported in parentheses. Traditional route includes candidates enrolled in university-based educator preparation programs; Alternative includes alternative certification providers; By Exam includes candidates who satisfied certification requirements by examination; Out of State includes candidates with credentials from another state. Sample includes first-attempt SOR exam takers from January 2021 to October 2023.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$



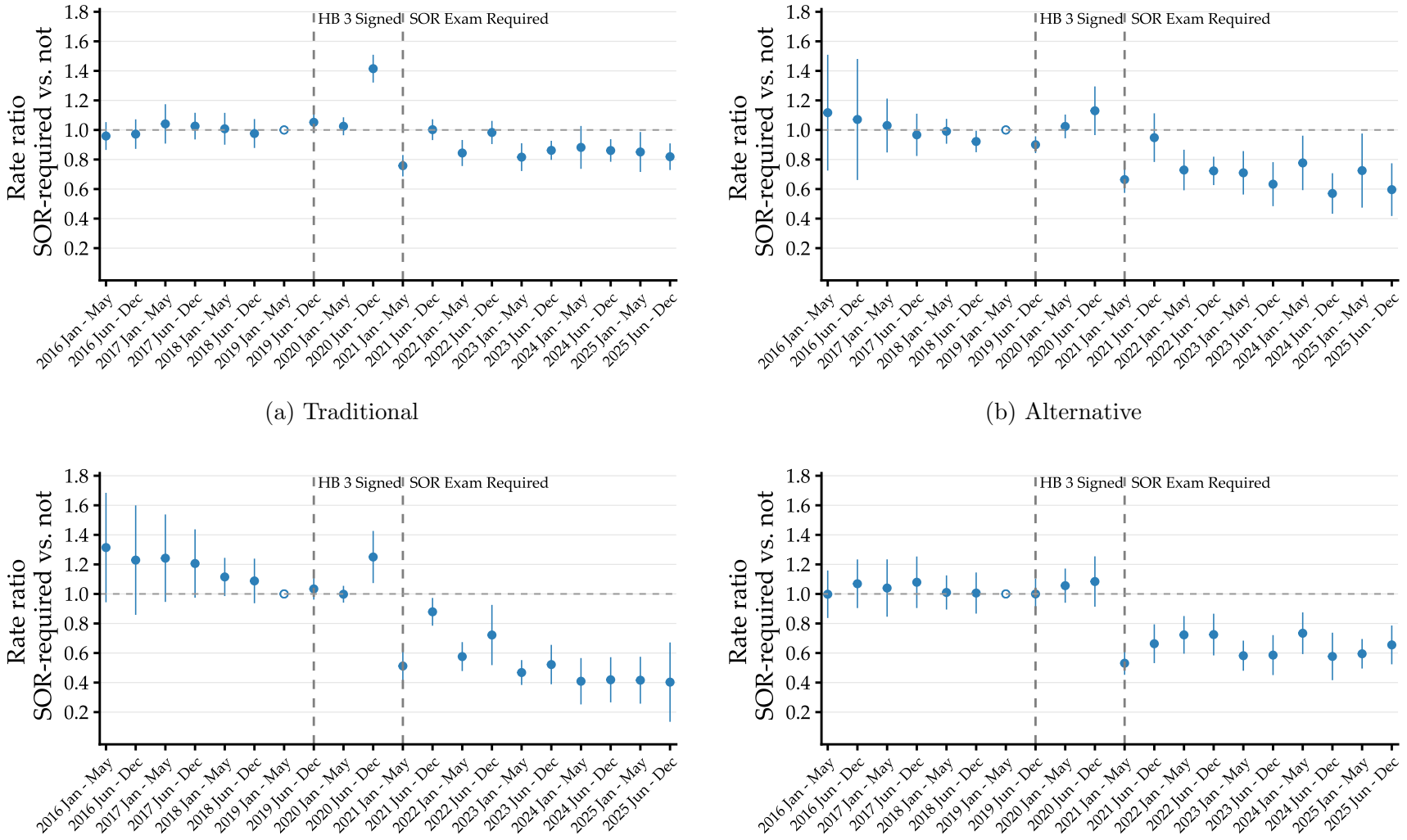
(a) Semi-annual certification counts



(b) Event study estimates

Figure 3: New Certifications in SOR-Required Fields

Notes: Panel (a) plots the number of newly issued standalone Texas teaching certifications requiring versus not requiring the TExES Science of Teaching Reading (SOR) exam (TExES 293), by half-year. The dotted line indicates when HB 3 was signed into law (June 2019); the dashed line indicates when the SOR requirement took effect (January 2021). Panel (b) plots coefficients from a difference-in-differences event study estimated by Poisson pseudo-maximum likelihood on data collapsed to the educator preparation program (EPP) × program type × half-year × certification type level, where the outcome is the count of newly issued certifications. Each point gives the rate ratio of certifications in SOR-required relative to non-SOR fields, normalized to the omitted period, 2019 January–May (hollow marker). Specifications include EPP × program type, certification type, and half-year fixed effects. Vertical bars give 95% confidence intervals; standard errors are clustered at the certification type level. Table A2 lists SOR-required and comparison fields. A newly issued certification refers to the first time an individual receives that certification; non-teaching and supplemental credentials are excluded.



(a) Traditional

(b) Alternative

(c) By Exam

(d) Out of State

Figure 4: New Certifications in SOR-Required Fields Relative to 2019H1, by Certification Route

Notes: Each panel reproduces the market-level difference-in-differences event study of Figure 3 for a single certification route: Traditional (panel a), Alternative (panel b), Certification by Exam (panel c), and Out of State (panel d). Within each route, coefficients are estimated by Poisson pseudo-maximum likelihood on data collapsed to the educator preparation program (EPP) × program type × half-year × certification type level, where the outcome is the count of newly issued certifications. Each point gives the rate ratio of certifications in SOR-required relative to non-SOR fields in that half-year, normalized to the omitted period, 2019 January–May (hollow marker). The dotted vertical line indicates when HB 3 was signed into law (June 2019); the dashed vertical line indicates when the SOR requirement took effect (January 2021). Specifications include EPP × program type, certification type, and half-year fixed effects. Vertical bars give 95% confidence intervals; standard errors are clustered at the certification type level. Table A2 lists SOR-required and comparison fields. A newly issued certification refers to the first time an individual receives that certification; non-teaching and supplemental credentials are excluded.

Table 7: Effect of the SOR Requirement on New Certifications

	All	Traditional	Alternative	By Exam	Out of State
Panel A. Event Study					
2016 Jan–May SOR Req.	1.093	0.962	1.112	1.306	1.004
	(0.102)	(0.047)	(0.195)	(0.188)	(0.096)
2016 Jun–Dec SOR Req.	1.105	0.984	1.069	1.217	1.069
	(0.144)	(0.058)	(0.208)	(0.188)	(0.090)
2017 Jan–May SOR Req.	1.075	1.043	1.035	1.234	1.049
	(0.064)	(0.066)	(0.091)	(0.147)	(0.107)
2017 Jun–Dec SOR Req.	1.077	1.030	0.970	1.199	1.098
	(0.064)	(0.048)	(0.071)	(0.115)	(0.100)
2018 Jan–May SOR Req.	1.032	1.011	0.995	1.106	1.022
	(0.032)	(0.055)	(0.042)	(0.063)	(0.079)
2018 Jun–Dec SOR Req.	1.038	0.987	0.921*	1.085	1.031
	(0.046)	(0.054)	(0.035)	(0.074)	(0.069)
2019 Jan–May SOR Req.	1.000	1.000	1.000	1.000	1.000
2019 Jun–Dec SOR Req.	1.049	1.056	0.900***	1.032	1.011
	(0.025)	(0.037)	(0.029)	(0.036)	(0.057)
2020 Jan–May SOR Req.	1.082**	1.029	1.023	0.997	1.077
	(0.030)	(0.031)	(0.042)	(0.028)	(0.058)
2020 Jun–Dec SOR Req.	1.365***	1.419***	1.131	1.247**	1.137
	(0.055)	(0.047)	(0.085)	(0.089)	(0.081)
2021 Jan–May SOR Req.	0.673***	0.758***	0.665***	0.519***	0.574***
	(0.043)	(0.036)	(0.049)	(0.052)	(0.044)
2021 Jun–Dec SOR Req.	0.988	1.003	0.947	0.875**	0.732***
	(0.037)	(0.035)	(0.084)	(0.042)	(0.059)
2022 Jan–May SOR Req.	0.773***	0.864***	0.730**	0.574***	0.844*
	(0.055)	(0.034)	(0.072)	(0.051)	(0.071)
2022 Jun–Dec SOR Req.	0.887*	0.994	0.723***	0.717*	0.830*
	(0.054)	(0.038)	(0.051)	(0.104)	(0.068)
2023 Jan–May SOR Req.	0.728***	0.840***	0.712**	0.467***	0.676***
	(0.050)	(0.040)	(0.077)	(0.042)	(0.050)
2023 Jun–Dec SOR Req.	0.767***	0.881***	0.633***	0.521***	0.682***
	(0.061)	(0.033)	(0.078)	(0.071)	(0.057)
2024 Jan–May SOR Req.	0.747*	0.914	0.776*	0.408***	0.869
	(0.085)	(0.065)	(0.095)	(0.079)	(0.076)
2024 Jun–Dec SOR Req.	0.662***	0.901**	0.567***	0.420***	0.679***
	(0.071)	(0.037)	(0.071)	(0.079)	(0.071)
2025 Jan–May SOR Req.	0.682**	0.883	0.730	0.417***	0.715***
	(0.091)	(0.066)	(0.131)	(0.081)	(0.056)
2025 Jun–Dec SOR Req.	0.567*	0.866**	0.590***	0.347**	0.736***
	(0.145)	(0.039)	(0.090)	(0.138)	(0.065)
N	60,705	9,607	9,637	380	428
Joint F-test of pre-period (p)	0.633	0.025	< 0.001	0.008	0.215
Panel B. Difference-in-Differences					
Post × SOR Req. (2019H2–2025H2)	0.785*	0.940	0.780	0.531***	0.765**
	(0.080)	(0.063)	(0.120)	(0.075)	(0.071)
RI p-value[a]	[0.081]	[0.344]	[0.05]	[< 0.001]	[0.005]
Post × SOR Req. (2022H1–2025H2)	0.702**	0.879*	0.712*	0.442***	0.790***
	(0.076)	(0.052)	(0.102)	(0.079)	(0.054)
RI p-value	[0.039]	[0.102]	[0.001]	[< 0.001]	[< 0.001]

Notes: Table reports estimates from Poisson pseudo-maximum likelihood regressions of newly issued standalone teaching certifications on an interaction between certification-field SOR-required status and time. All specifications include EPP × program-type, certification-field, and half-year fixed effects. Panel A reports the event-study specification, with coefficients normalized to the omitted reference period, 2019 January–May. Each point-estimate row gives the rate ratio of certifications in SOR-required relative to non-SOR fields in that half-year. Panel B reports two companion difference-in-differences specifications. Column (1) uses the full sample; columns (2)–(5) restrict to the traditional, alternative, certification-by-examination, and out-of-state routes, respectively. Coefficients are exponentiated and reported as rate ratios. Standard errors clustered at the certification-field level are provided in parenthesis. In Panel B, the clustered-SE p-value corresponding to each coefficient is reported in brackets directly beneath it, alongside the randomization inference p-value for the same coefficient.

[a]Randomization inference p-value is calculated by re-estimating the specification 2,000 times, each time randomly assigning SOR-required status to 3 of the 19 certification fields while holding the true number of treated fields fixed, and computing the t-statistic on the coefficient of interest in each placebo estimate. The reported p-value is the share of these placebo t-statistics whose magnitude equals or exceeds that of the t-statistic from the true assignment.

[b]The first difference-in-differences row uses the full post-announcement period (half-years from 2019 June onward). The second row reports a donut specification that excludes half-years 2019H2 through 2021H2.

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 8: Predicting Student Achievement with Teacher Certification Exam Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Reading									
SOR Exam	0.011***	0.013***	0.013***				0.011**	0.009*	0.012***
	(0.002)	(0.002)	(0.002)				(0.004)	(0.003)	(0.004)
PPR Exam				0.003	0.005***	0.005***	−0.003	−0.001	−0.006
				(0.002)	(0.001)	(0.001)	(0.004)	(0.003)	(0.004)
N	829,964	829,964	829,964	1,440,994	1,440,994	1,440,994	395,669	395,669	395,669
N Teachers	8,301	8,301	8,301	14,576	14,576	14,576	4,987	4,987	4,987
Panel B. Math									
SOR Exam	0.015***	0.017***	0.014***				0.020***	0.017***	0.016**
	(0.003)	(0.003)	(0.003)				(0.006)	(0.005)	(0.005)
PPR Exam				0.006*	0.007***	0.010***	−0.003	−0.002	0.007
				(0.002)	(0.002)	(0.002)	(0.006)	(0.005)	(0.005)
N	808,593	808,593	808,593	1,367,248	1,367,248	1,367,248	385,255	385,255	385,255
N Teachers	8,292	8,292	8,292	14,447	14,447	14,447	4,979	4,979	4,979
Grade FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
District FE		✓			✓			✓	
Campus FE			✓			✓			✓

Notes: Table reports estimates from value-added models estimated via ordinary least squares (OLS) relating teachers' standardized certification exam scores to student achievement on the State of Texas Assessments of Academic Readiness (STAAR). Panel A reports student reading achievement, and Panel B reports student math achievement. The dependent variable is the corresponding student test score standardized to have mean zero and standard deviation one within grade, year, and subject. Each column reports a separate regression. Teacher certification exam scores are standardized to have mean zero and standard deviation one among test takers. Columns (1)–(3) estimate the association between Science of Reading (SOR) exam scores and student achievement, columns (4)–(6) use Pedagogy and Professional Responsibilities (PPR) exam scores, and columns (7)–(9) enter both scores jointly, so that each coefficient reflects the association net of the other. All specifications include controls for prior reading and math achievement (quadratic and cubic terms), race, gender, special education status, gifted status, economic disadvantage, and classroom- and school-level averages of these covariates, along with grade and year fixed effects. Columns (2), (5), and (8) add district fixed effects, and columns (3), (6), and (9) add campus fixed effects; the fixed-effects rows are reported once, in Panel B, and apply to both panels. Standard errors clustered at the teacher level are reported in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

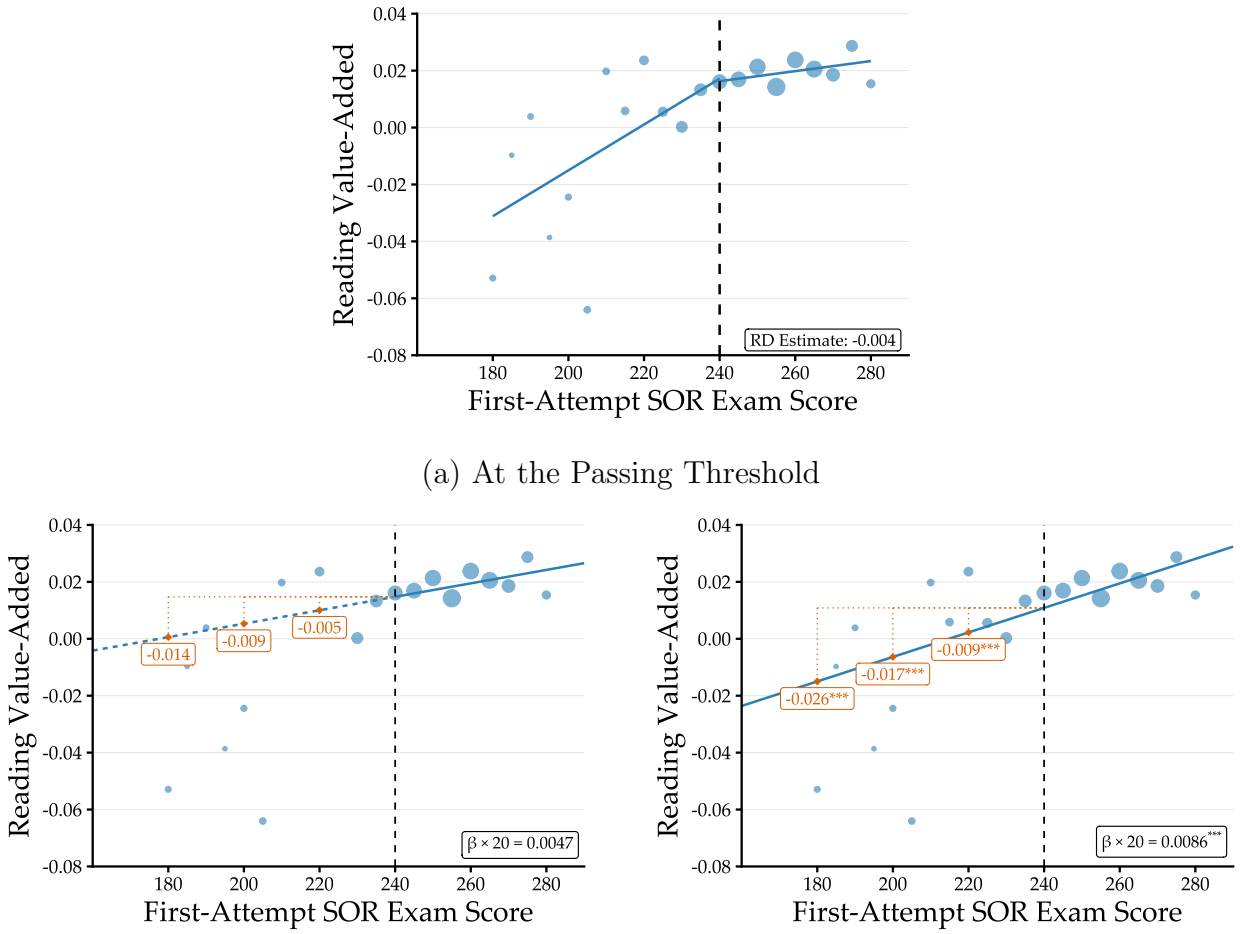
Table 9: Discontinuity in Teacher Value-Added

	Reading (1)	Math (2)
Original estimate	−0.004	0.029
	(0.019)	(0.031)
Lee upper bound[a]	0.033	0.090**
	(0.024)	(0.037)
Lee lower bound	−0.013	−0.022
	(0.024)	(0.036)
Trim share	10.1%	10.4%

Notes: Table reports regression discontinuity estimates evaluating whether there is a discontinuity in observed teacher value-added at the passing threshold of the Science of Teaching Reading (SOR) exam (TExES 293). The outcome is the teacher's estimated value-added in reading (Column 1) or math (Column 2), using the first-attempt SOR score as the running variable and CCT optimal bandwidth (Calonico et al., 2014). Original estimate reports the local-constant regression discontinuity point estimate and heteroskedasticity-robust standard error.

[a] *Lee bounds*: because passing raises the chance a candidate is later observed teaching with a value-added score, the passers we observe are a less selected group than the failers we observe. Trim share measures how much less selected: we run a regression discontinuity of whether value-added is observed on exam score, and divide the resulting jump in observation rates at the passing threshold by the observation rate among passers, giving the share of observed passers who would not have been observed had they instead failed. To construct the bounds, we take the observed passers, rank them by value-added, and drop the Trim share with the lowest value-added; re-estimating the discontinuity on this trimmed sample gives the upper bound. This corresponds to the assumption that every candidate observed only because they passed has the lowest value-added among passers, the least favorable case for the passing group and so the largest possible discontinuity. We then instead drop the Trim share with the highest value-added and re-estimate; this gives the lower bound. All specifications control for month-of-exam fixed effects.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$



(a) At the Passing Threshold

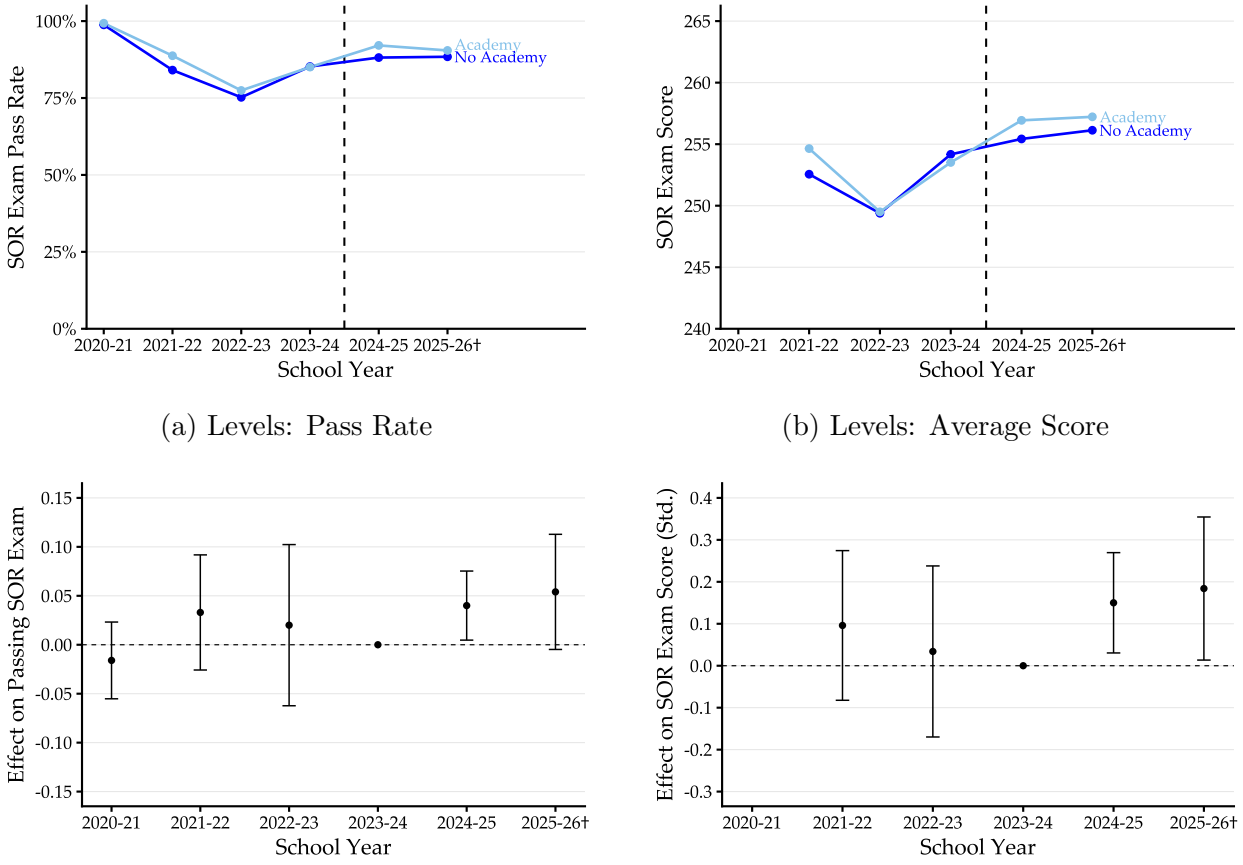
(b) Below the Threshold: Passers Only

(c) Below the Threshold: Full Sample

Figure 5: Teacher Value-Added at the Passing Threshold and Below It

Notes: All panels plot reading value-added against first-attempt SOR exam score, binned in 5-point intervals, with the passing threshold (score = 240) marked by a dashed vertical line. Point size is proportional to the number of teachers in each bin. Panel (a) plots separate local linear fits on either side of the threshold, with the annotated RD estimate from Equation 1, estimated with teacher value-added as the outcome. See Table 9 for additional information. Panels (b) and (c) plot a single linear OLS fit of value-added on score, with robust standard errors, estimated separately on (b) teachers who passed the exam and (c) all teachers who attempted it. The solid segment of each line spans the range of scores used to estimate it, while the dashed segment in panel (b) is extrapolated beyond that range. $\beta \times 20$ reports the fitted line's slope over a 20-point increase in score. Orange labels report the difference between the predicted value-added at each threshold (220, 200, 180) and at the current threshold (240).

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.



(a) Levels: Pass Rate

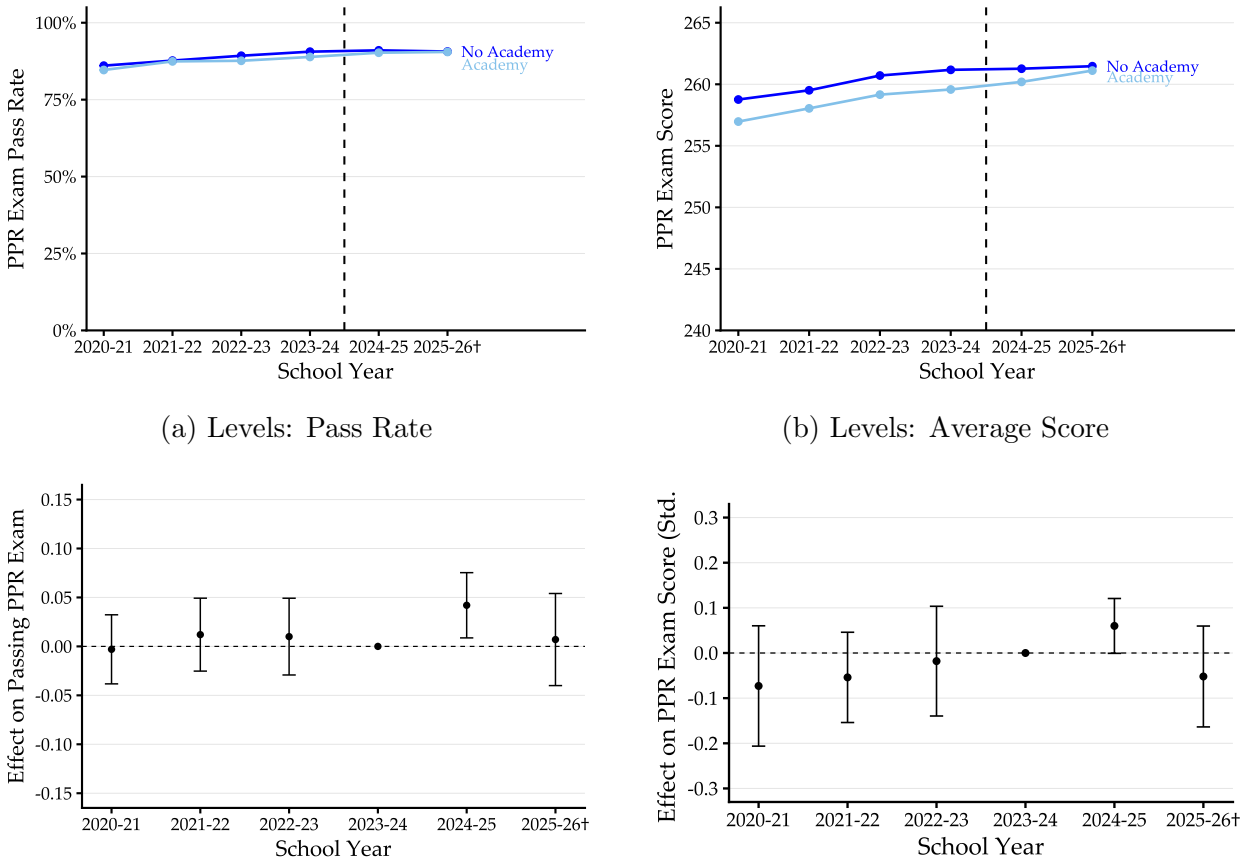
(b) Levels: Average Score

(c) Event Study: Pass Rate

(d) Event Study: Average Score

Figure 6: TExES Science of Reading (SOR) Outcomes by SOR Academy

Notes: Panels (a) and (b) show levels of pass rates and average scores, respectively, for the Science of Reading (SOR) exam, comparing EPPs with and without SOR academies. Panels (c) and (d) show event study estimates for the same outcomes, with the reference year (2023–24) indicated by a hollow marker. SOR academies were first implemented in the 2024–25 school year. Pass rate is measured as the proportion of first-attempt test takers who pass the exam (score $\geq$ 240). Average score represents the mean exam score among all first-attempt test takers, standardized to have mean zero and standard deviation of one within each year. † Certification exam data through March 2026.



(a) Levels: Pass Rate

(b) Levels: Average Score

(c) Event Study: Pass Rate

(d) Event Study: Average Score

Figure 7: Pedagogy and Professional Responsibilities (PPR) Outcomes by SOR Academy

Notes: Panels (a) and (b) show levels of pass rates and average scores, respectively, for the Pedagogy and Professional Responsibilities (PPR) exam, comparing EPPs with and without SOR academies. Panels (c) and (d) show event study estimates for the same outcomes, with the reference year (2023–24) indicated by a hollow marker. SOR academies were first implemented in the 2024–25 school year. Pass rate is measured as the proportion of first-attempt test takers who pass the PPR exam. Average score represents the mean exam score among all first-attempt test takers, standardized to have mean zero and standard deviation of one within each year. 2020–21 scaled scores are unavailable due to piloting. † Certification exam data through March 2026.

Table 10: Effect of SOR Academies on Certification Exam Outcomes

	SOR		PPR	
	Pass Rate	Score	Pass Rate	Score
	(1)	(2)	(3)	(4)
Panel A. Event Study				
2020–21 × SOR Academy	−0.016		−0.003	−0.073
	(0.020)		(0.018)	(0.068)
2021–22 × SOR Academy	0.033	0.096	0.012	−0.054
	(0.030)	(0.091)	(0.019)	(0.051)
2022–23 × SOR Academy	0.020	0.034	0.010	−0.018
	(0.042)	(0.104)	(0.020)	(0.062)
2023–24 × SOR Academy	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)
2024–25 × SOR Academy	0.040*	0.150*	0.042*	0.060+
	(0.018)	(0.061)	(0.017)	(0.031)
2025–26 × SOR Academy†	0.054+	0.184*	0.007	−0.052
	(0.030)	(0.087)	(0.024)	(0.057)
Panel B. Difference-in-Differences				
Post × SOR Academy	0.038+	0.119	0.027	0.061
	(0.023)	(0.096)	(0.019)	(0.046)
Joint F-test of pre-period	0.529	0.544	0.638	0.642
N	61,446	40,756	95,617	95,617
Pre-period mean, non-Academy	0.856	0.029	0.854	-0.004
Pre-period mean, Academy	0.861	0.007	0.851	-0.015

Notes: Table reports estimates from event study (Panel A) and difference-in-differences (Panel B) specifications comparing certification exam outcomes at Educator Preparation Programs (EPPs) that adopted Science of Reading (SOR) Academies in 2024–25 to EPPs that never adopted. The outcomes are the first-attempt pass rate and standardized average score on the SOR exam (the TExES Science of Teaching Reading exam, exam 293) and the Pedagogy and Professional Responsibilities (PPR) exam, measured at the EPP-year level. All specifications include EPP fixed effects and year fixed effects. The reference year for the event study is 2023–24, the year immediately preceding academy implementation. The joint F-test reports the p-value from a test that all pre-period interaction coefficients are jointly zero. EPPs that adopted SOR Academies in 2025–26 are excluded from both groups. † Certification exam data through March 1, 2026. Standard errors clustered at the Educator Preparation Program level in parentheses. 2020–21 scaled scores are unavailable due to piloting.
$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Appendix A. Additional Tables and Figures

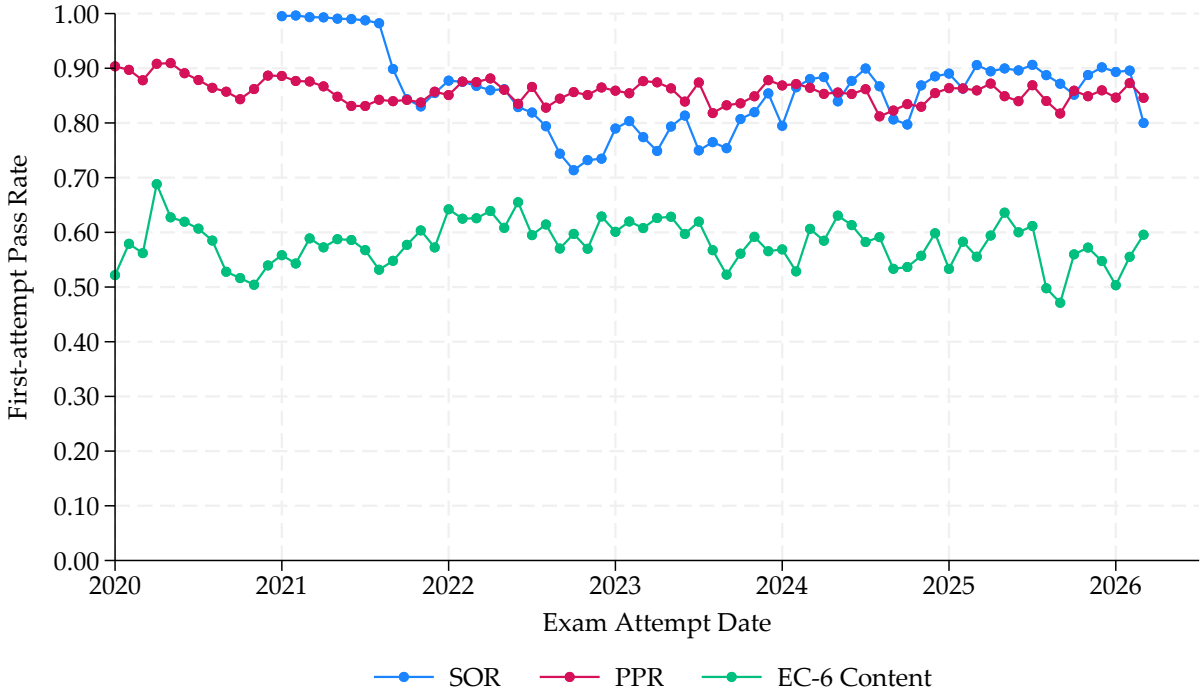


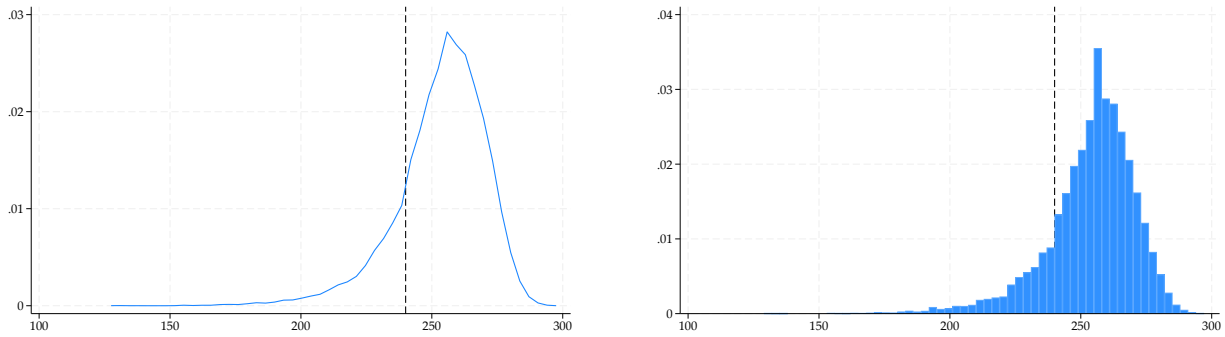
Figure A1: First-Attempt Pass Rates for Texas Teacher EC-6 Certification Exams

Notes: This figure plots first-attempt pass rates for three Texas teacher certification exams from 2020 to 2025. The blue line shows pass rates for the Science of Reading (SOR) exam (exam 293). The orange line shows pass rates for the Pedagogy and Professional Responsibilities (PPR) exam. The green line shows pass rates for the EC-6 Content exam. The sample includes all first-attempt test takers in Texas. The black vertical line indicates January 1, 2021, when the SOR exam was introduced. The red dashed vertical line indicates when the State Board for Educator Certification (SBEC) began including scaled scores from the SOR exam after initial piloting.

Table A1: Covariates at the SOR Exam Passing Threshold

	Conventional	Bias-corrected	Robust
Female	0.00574	0.0122	0.0122
	(0.0202)	(0.0202)	(0.0241)
White	0.0326	0.0385	0.0385
	(0.0321)	(0.0321)	(0.0385)
Traditional	0.000748	−0.00555	−0.00555
	(0.0218)	(0.0218)	(0.0253)
Alternative	0.0112	0.00795	0.00795
	(0.0269)	(0.0269)	(0.0323)
By Exam	0.0106	0.0168	0.0168
	(0.0166)	(0.0166)	(0.0194)
Out of State	−0.0168	−0.0209	−0.0209
	(0.0160)	(0.0160)	(0.0189)
Teaches Before Exam	−0.00896	−0.00950	−0.00950
	(0.0262)	(0.0262)	(0.0320)

Notes: Table reports regression discontinuity estimates at the SOR exam passing threshold (score ≥ 240) with each pre-determined covariate as the outcome. None of the estimates are statistically significant at the 0.05 level, supporting the assumption that candidates just above and below the cutoff are comparable in observable characteristics. All estimates use a local linear regression with triangular kernel weights and data-driven bandwidth selection. Heteroskedasticity-robust standard errors are reported in parentheses. Demographic variables (i.e., Female, White) are available only for candidates linked to PEIMS or THECB records.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$



(a) Density Plot

(b) Histogram

Figure A2: Distribution of Science of Reading Exam Scores

Notes: This figure plots the distribution of scores on the Texas Science of Reading (SOR) exam (exam 293) for first-attempt test takers between January 1, 2021 and October 23, 2023. Panel (a) shows a kernel density estimate of the score distribution. Panel (b) shows a histogram with overlaid density estimate. The vertical dotted line indicates the passing threshold of 240. A density discontinuity test using the method of Cattaneo et al. (2018) fails to reject the null hypothesis of no manipulation around the cutoff ($p = 0.051$), suggesting no evidence of sorting around the passing threshold.

Table A2: Classification of Texas Teaching Certificates by SOR-Exam Requirement

Certification Group	SOR (293) Required?
Panel A: SOR-required certifications	
Core Subjects: Early Childhood–Grade 6	Yes
Core Subjects: Grades 4–8	Yes
English Language Arts and Reading: Grades 4–8	Yes
Panel B: Non-SOR certifications (comparison group)	
Computer Science: Grades 8–12	No
English Language Arts and Reading: Grades 7–12	No
Fine Arts: EC–12	No
Foreign Language: EC–12	No
Health & Physical Education	No
Mathematics: Grades 4–8	No
Mathematics: Grades 7–12	No
Science: Grades 4–8	No
Science: Grades 6–12	No
Science: Grades 7–12	No
Social Studies: Grades 4–8	No
Social Studies: Grades 7–12	No
Special Education: EC–12	No
Vocational Education 4–8	No
Vocational Education 7–12	No
Vocational Education EC–12	No

Notes: Table lists the 19 certification groups in the market-level analysis sample, which are also the units at which standard errors are clustered. Classification follows the Texas Education Agency Required Test Chart (May 2024); a certification is coded SOR-required if the TExES Science of Teaching Reading exam (293) appears among its required content-pedagogy tests. Non-teaching credentials (Reading Specialist, Counselor, Diagnostician, Librarian, Principal, Superintendent) and supplemental certificates that attach to a base certificate rather than certifying a teaching field independently (Bilingual, ESL, Gifted & Talented, Special Education Supplemental) are excluded from the sample.

Table A3: RD Estimates of Passing the Science of Reading Exam: With and Without Controls

	Retake SOR Exam		Earn Std. Cert. in 6 Months		Earn Std. Cert. in 1 Year		Teach Next October	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Conventional	-0.777^{***}	-0.776^{***}	0.0932^{***}	0.0922^{***}	0.0592^{*}	0.0596^{*}	0.0640^{*}	0.0581^{*}
	(0.0191)	(0.0193)	(0.0242)	(0.0246)	(0.0266)	(0.0267)	(0.0250)	(0.0258)
Bias-corrected	-0.779^{***}	-0.778^{***}	0.0923^{***}	0.0913^{***}	0.0485	0.0494	0.0650^{**}	0.0598^{*}
	(0.0191)	(0.0193)	(0.0242)	(0.0246)	(0.0266)	(0.0267)	(0.0250)	(0.0258)
Robust	-0.779^{***}	-0.778^{***}	0.0923^{**}	0.0913^{**}	0.0485	0.0494	0.0650^{*}	0.0598
	(0.0230)	(0.0233)	(0.0293)	(0.0297)	(0.0301)	(0.0303)	(0.0304)	(0.0314)
Month of Exam FE	✓		✓		✓		✓	
N	21,692	21,692	21,692	21,692	21,692	21,692	21,692	21,692

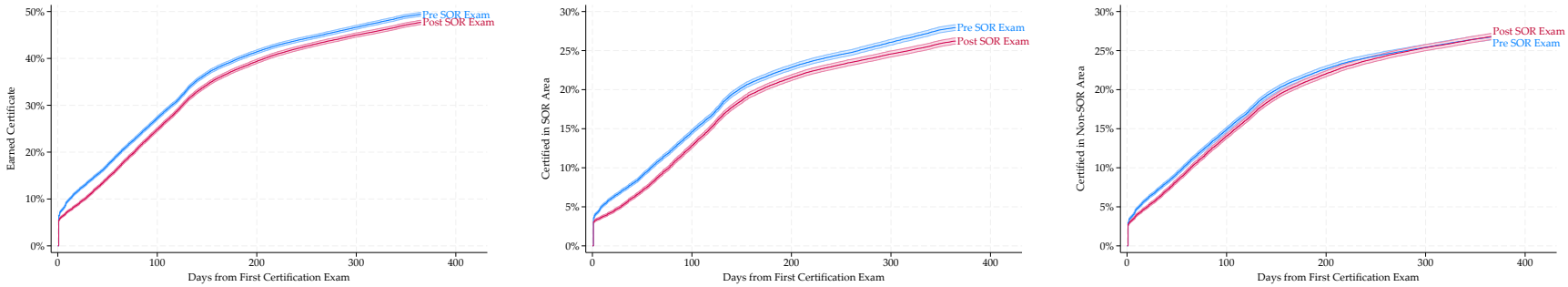
Notes: Table reports regression discontinuity estimates of the effect of passing the Science of Teaching Reading (SOR) exam (TExES 293) on teacher outcomes with and without month-of-exam fixed effects. Odd-numbered columns include month-of-exam fixed effects; even-numbered columns do not. Columns (1)–(2) show the probability of retaking the SOR exam within one year of the exam date. Columns (3)–(4) show the probability of earning a Standard teaching certificate within six months of the exam date. Columns (5)–(6) show the probability of earning a Standard teaching certificate within one year of the exam date. Columns (7)–(8) show the probability of working as a teacher in Texas public schools on the last Friday of October following the exam date. Month-of-exam fixed effects account for the mechanical relationship between exam timing and the annual October employment snapshot. All estimates use a local linear regression with triangular kernel weights and data-driven bandwidth selection. Standard errors in parentheses. Conventional and bias-corrected rows report heteroskedasticity-robust standard errors; robust rows report the bias-corrected robust standard errors of Calonico et al. (2014), which account for bias in the local polynomial estimator near the boundary. Sample includes all first-attempt test takers of the SOR exam from January 2021 to October 2023.

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table A4: Robustness of RD Estimates to Bandwidth Selection Method

	Retake SOR Exam			Earn Std. Cert. 6 Mo.			Earn Std. Cert. 1 Yr.			Teach Next October		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Conventional	-0.777^{***}	-0.777^{***}	-0.772^{***}	0.0932^{***}	0.0922^{***}	0.0769^{*}	0.0592^{*}	0.0593^{*}	0.0307	0.0640^{*}	0.0639^{*}	0.0640^{*}
	(0.0191)	(0.0192)	(0.0252)	(0.0242)	(0.0253)	(0.0316)	(0.0266)	(0.0266)	(0.0347)	(0.0250)	(0.0250)	(0.0327)
Bias-corrected	-0.779^{***}	-0.779^{***}	-0.773^{***}	0.0923^{***}	0.0924^{***}	0.0772^{*}	0.0485	0.0492	0.0263	0.0650^{**}	0.0652^{**}	0.0647^{*}
	(0.0191)	(0.0192)	(0.0252)	(0.0242)	(0.0253)	(0.0316)	(0.0266)	(0.0266)	(0.0347)	(0.0250)	(0.0250)	(0.0327)
Robust	-0.779^{***}	-0.779^{***}	-0.773^{***}	0.0923^{**}	0.0924^{**}	0.0772^{*}	0.0485	0.0492	0.0263	0.0650^{*}	0.0652^{*}	0.0647
	(0.0230)	(0.0231)	(0.0269)	(0.0293)	(0.0302)	(0.0338)	(0.0301)	(0.0320)	(0.0362)	(0.0304)	(0.0298)	(0.0351)
BW Method	mserd	msesum	cerrd	mserd	cerrd	certwo	mserd	msesum	cerrd	mserd	msesum	cerrd
Bandwidth	13.4	13.3	8.1	13.2	12.0	8.0	13.0	13.1	7.9	12.3	12.3	7.5
Month of Exam FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	21,692	21,692	21,692	21,692	21,692	21,692	21,692	21,692	21,692	21,643	21,643	21,643

Notes: Table reports regression discontinuity estimates of the effect of passing the Science of Teaching Reading (SOR) exam under different data-driven bandwidth selection methods from Calonico et al. (2014). Each set of three columns presents a separate outcome estimated using three bandwidth methods: mserd (MSE-optimal, common bandwidth), msesum (MSE-optimal, sum of bandwidths), cerrd (CER-optimal, common bandwidth), and certwo (CER-optimal, two bandwidths). MSE-optimal bandwidths minimize mean squared error and tend to be wider; CER-optimal bandwidths ensure coverage error rate optimality and tend to be narrower. Employment specifications include month-of-exam fixed effects. Standard errors in parentheses. Conventional and bias-corrected rows report heteroskedasticity-robust standard errors; robust rows report the bias-corrected robust standard errors of Calonico et al. (2014), which account for bias in the local polynomial estimator near the boundary. Sample includes all first-attempt test takers of the SOR exam from January 2021 to October 2023.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$



(a) Any Certification (b) Certs. Requiring SOR Exam (c) Certs. Not Requiring SOR Exam

Figure A3: Cumulative Certification Rates by Days Since First Certification Exam

Notes: Each panel plots the cumulative proportion of teacher candidates who earn a teaching certificate within a given number of days of sitting for their first Texas certification exam of any kind. The x-axis measures days elapsed since the candidate's first certification exam attempt. The y-axis in panel (a) measures the proportion who earn any teaching certification; the y-axis in panel (b) measures the proportion who earn a certification in an area requiring the TExES Science of Reading (SOR) exam (exam 293); the y-axis in panel (c) measures the proportion who earn a certification in an area not requiring the SOR exam. Panels (b) and (c) are not mutually exclusive, as candidates may earn both types of certification. Shaded regions represent 95% confidence intervals. The red line (Post SOR Exam) includes candidates whose first certification exam falls between January 1, 2021 and October 23, 2023, ensuring that all candidates are observed for at least 365 days. The blue line (Pre SOR Exam) includes candidates whose first certification exam falls between January 1, 2016 and January 1, 2019. Panel (c) serves as a placebo test: because candidates pursuing non-SOR certifications faced the same macroeconomic and institutional environment as those in the post-SOR period, divergence in panel (b) but not panel (c) isolates the SOR mandate as the operative factor. Certifications requiring the SOR exam are listed in Table A2.

Table A5: Sensitivity of the Market-Level Estimate to Pre-Trend Violations

	2019 Jun-Dec to 2025 Jun-Dec		Donut-weighted: 2022 Jan-Jun to 2025 Jun-Dec	
	90% CI	95% CI	90% CI	95% CI
Original (unrestricted)	[0.730, 0.921]	[0.714, 0.941]	[0.614, 0.848]	[0.595, 0.874]
$M = 0.000$	[0.790, 0.997]	[0.772, 1.020]	[0.690, 0.937]	[0.670, 0.965]
$M = 0.001$	[0.773, 1.031]	[0.753, 1.057]	[0.668, 0.985]	[0.645, 1.017]
$M = 0.002$	[0.744, 1.093]	[0.722, 1.118]	[0.630, 1.069]	[0.606, 1.102]

Notes: Table reports HonestDiD confidence intervals, exponentiated and expressed as rate ratios, under the smoothness restriction Δ^{SD} (Rambachan & Roth, 2023) for the equally-weighted average post-mandate effect over two target windows: the full post-announcement period (2019H2–2025H2, including the anticipatory certification surge preceding the mandate's January 2021 implementation) and the post-implementation period (2022H1–2025H2). The donut-weighted column places zero weight on half-years 2019H2–2021H2 in the target average, rather than excluding them from the underlying event-study estimation. This allows the counterfactual pre-trend to run smoothly through the anticipatory and initial implementation periods before being evaluated against the post-implementation coefficients, and is more conservative than dropping these periods outright, since it lets any accumulated pre-trend deviation compound over more periods before reaching the target window. M is the maximum permitted period-to-period change in the counterfactual trend's slope, relative to a linear extrapolation of the pre-period trend; $M = 0$ imposes exact linear continuation.

Table A6: SOR Exam Performance and Characteristics of Students at First School

	Econ. Disadv.		LEP		SPED		Gifted		ESL	
	OLS	RD	OLS	RD	OLS	RD	OLS	RD	OLS	RD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Estimate	-0.046^{***}	-0.009	-0.018^{***}	0.001	0.002^{***}	0.001	0.002^{***}	0.006	-0.002^{*}	-0.002
	(0.002)	(0.019)	(0.002)	(0.013)	(0.001)	(0.003)	(0.001)	(0.004)	(0.001)	(0.006)

Notes: Table reports the relationship between Science of Teaching Reading (SOR) exam performance and the characteristics of the students a candidate teaches in the first year following the exam. Each cell reports the estimate from a separate regression. Outcomes are the share of students at the candidate's school who are economically disadvantaged (1)–(2), limited English proficient (3)–(4), receiving special education services (5)–(6), identified as gifted and talented (7)–(8), and classified as English as a Second Language (9)–(10). Odd-numbered columns report OLS estimates of the association between a one standard deviation increase in the SOR exam score and each outcome. Even-numbered columns report regression discontinuity estimates of the effect of passing the SOR exam at the certification threshold, using local linear regression with triangular kernel weights and data-driven bandwidth selection; estimates are bias-corrected and standard errors are the robust standard errors of Calonico et al. (2014). All specifications include month-of-exam fixed effects. Standard errors reported in parentheses. Because student characteristics are observed only for candidates who go on to teach, these estimates are necessarily conditional on employment as a teacher in a Texas public school in the year following the exam, and the sample is correspondingly smaller than in the certification tables. Sample includes first-attempt SOR exam takers from January 2021 to October 2023.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table A7: RD Estimates of Passing the SOR Exam on Certification, by Certification Field

	(1)	(2)	(3)	(4)
	Earn Std. Cert. in 6 Months		Earn Std. Cert. in 1 Year	
	SOR Field	Non-SOR Field	SOR Field	Non-SOR Field
Conventional	0.112***	0.008	0.095***	−0.005
	(0.025)	(0.017)	(0.027)	(0.019)
Bias-corrected	0.106***	0.01	0.085**	−0.006
	(0.025)	(0.017)	(0.027)	(0.019)
Robust	0.106***	0.01	0.085**	−0.006
	(0.030)	(0.021)	(0.032)	(0.023)

Notes: Table reports regression discontinuity estimates of the effect of passing the Science of Teaching Reading (SOR) exam (TExES 293) on the probability of earning a Standard teaching certificate, estimated separately by certification field. Each cell reports the estimate from a separate regression. Columns (1) and (3) use as the outcome earning a certificate in a field for which the SOR exam is required; columns (2) and (4) use earning a certificate in a field for which it is not. SOR-required fields comprise early childhood and elementary generalist certifications (Core Subjects: Early Childhood–Grade 6), middle-grades generalist certifications (Core Subjects: Grades 4–8), and middle-grades English language arts and social studies certifications (English Language Arts and Reading: Grades 4–8; and English Language Arts, Reading/Social Studies: Grades 4–8). Because passing the SOR exam is a prerequisite only in the former group, the non-SOR columns serve as a falsification test. Columns (1) and (2) show the probability of earning a certificate within six months of the exam date; columns (3) and (4) within one year. All specifications include month-of-exam fixed effects. Estimates use local linear regression with triangular kernel weights and data-driven bandwidth selection. Conventional, bias-corrected, and robust estimates follow Calonico et al. (2014); standard errors reported in parentheses. Sample includes first-attempt SOR exam takers from January 2021 to October 2023.

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table A8: EPPs Offering Texas Reading Academies by Year of Adoption

Educator Preparation Program
2024–25 Academic Year
ESC19 Teacher Preparation & Certification Program
Lamar University
Teach Us Texas
Texas A&M University Commerce
Texas Christian University
Texas Tech University
Trinity University
University of Texas at El Paso (UTEP)
University of Texas at Tyler
West Texas A&M University
2025–26 Academic Year
240 Certification
ACTRGV/Pharr San Juan Alamo
Baylor University
Dallas College
East Texas Baptist
East Texas A&M University
Houston CC
LeTourneau University
McMurry University
Region 4 Inspire Texas
Region 19
Responsive Ed 180 Educator Preparation Program
Stephen F. Austin State University Department of Education Studies
Texas Woman's University EPP
The University of Texas at Dallas - Teacher Development Center

Notes: Table lists Educator Preparation Programs (EPPs) authorized to offer Texas Reading Academies (TRA) coursework to preservice elementary teachers by year of adoption. TRA is a statewide professional development program that engages elementary educators with high-quality instructional strategies rooted in the Science of Reading (SOR). EPPs that adopted TRA in 2024–25 and continued in 2025–26 are listed only under 2024–25. Source: Texas Education Agency.

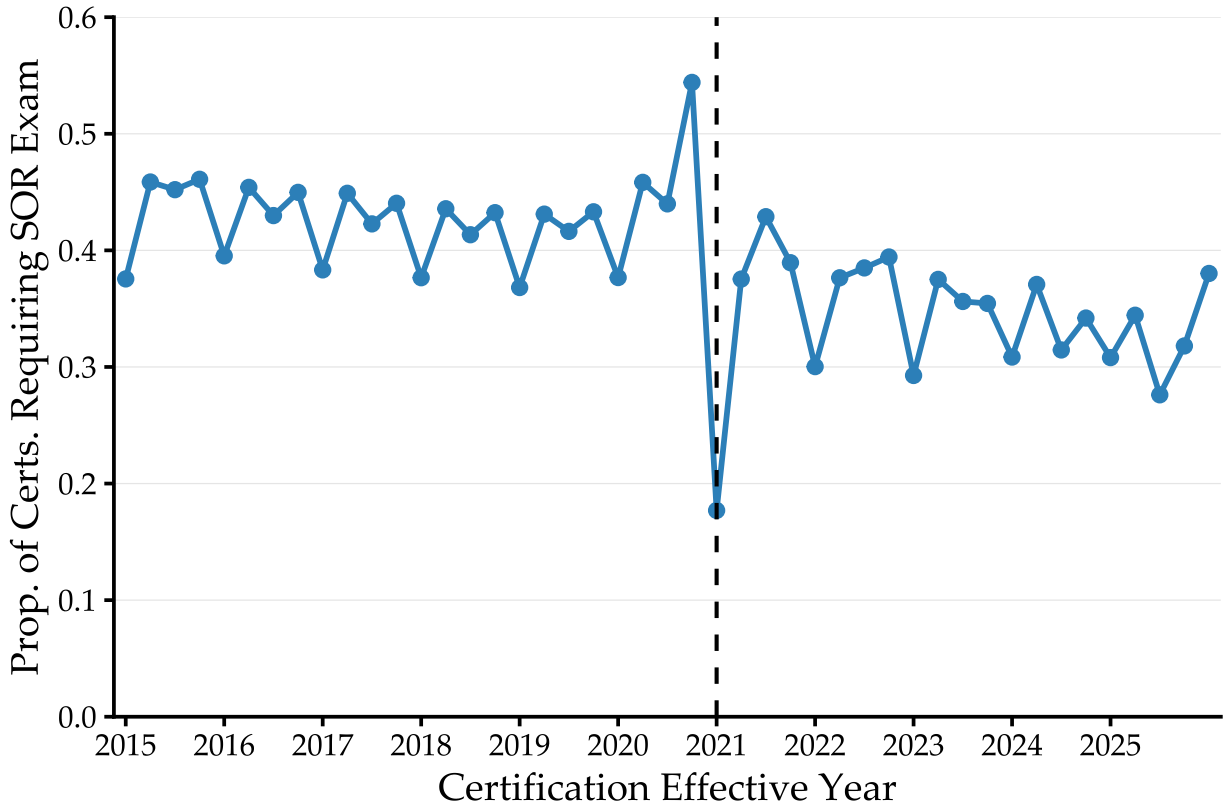
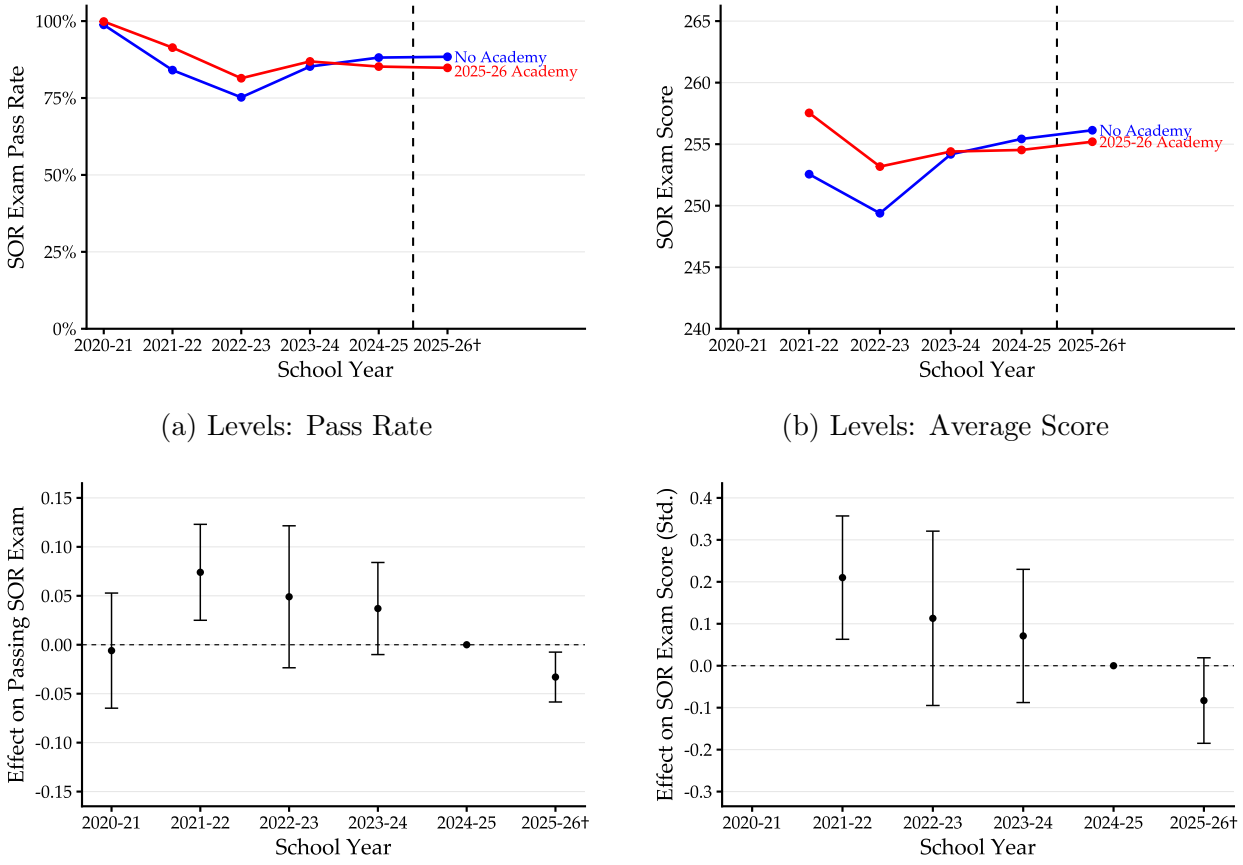


Figure A4: Proportion of New Teacher Certifications Requiring the Science of Reading Exam

Notes: The figure displays the proportion of newly issued Texas teacher certifications that require the TExES Science of Reading exam (exam 293) by quarter from 2018 through early 2025. Each point represents the quarterly average proportion across all certification types. The vertical dashed line indicates January 1, 2021, when the SOR exam requirement was implemented for certain certification types. A newly issued certification refers to the first time an individual receives that particular certification. The sample includes all teacher certification types issued in Texas during this period.



(a) Levels: Pass Rate

(b) Levels: Average Score

(c) Event Study: Pass Rate

(d) Event Study: Average Score

Figure A5: TExES Science of Reading (SOR) Outcomes by SOR Academy, 2025-26 Cohort

Notes: Panels (a) and (b) show levels of pass rates and average scores, respectively, for the Science of Reading (SOR) exam, comparing EPPs with and without SOR academies. Panels (c) and (d) show event study estimates for the same outcomes, with the reference year (2024–25) indicated by a hollow marker. SOR academies were first implemented in the 2025–26 school year. Pass rate is measured as the proportion of first-attempt test takers who pass the exam (score $\geq$ 240). Average score represents the mean exam score among all first-attempt test takers, standardized to have mean zero and standard deviation of one within each year. † Certification exam data through March 2026.

Table A9: Effect of SOR Academies on Certification Exam Outcomes, 2025-26 Cohort

	SOR		PPR	
	Pass Rate	Score	Pass Rate	Score
	(1)	(2)	(3)	(4)
Panel A. Event Study				
2020–21 × SOR Academy	0.006		−0.043	−0.069
	(0.033)		(0.028)	(0.084)
2021–22 × SOR Academy	0.083**	0.256**	−0.043	−0.082
	(0.028)	(0.08)	(0.032)	(0.087)
2022–23 × SOR Academy	0.057	0.145	−0.038+	−0.04
	(0.038)	(0.109)	(0.022)	(0.048)
2023–24 × SOR Academy	0.042	0.084	−0.056*	−0.019
	(0.025)	(0.081)	(0.023)	(0.039)
2024–25 × SOR Academy	0	0	0	0
	(.)	(.)	(.)	(.)
2025–26 × SOR Academy†	−0.025+	−0.06	−0.043**	−0.117**
	(0.013)	(0.055)	(0.015)	(0.043)
Panel B. Difference-in-Differences				
Post × SOR Academy	−0.052*	−0.146*	−0.007	−0.075
	(0.021)	(0.062)	(0.017)	(0.049)
Joint F-test of pre-period	0	0	0.359	0.713
N	61,423	40,842	97,389	97,389
Pre-period mean, non-Academy	0.856	0.051	0.854	-0.184
Pre-period mean, Academy	0.865	0.058	0.815	-0.272

Notes: Table reports estimates from event study (Panel A) and difference-in-differences (Panel B) specifications comparing certification exam outcomes at Educator Preparation Programs (EPPs) that adopted Science of Reading (SOR) Academies in 2025–26 to EPPs that never adopted. The outcomes are the first-attempt pass rate and standardized average score on the SOR exam (the TExES Science of Teaching Reading exam, exam 293) and the Pedagogy and Professional Responsibilities (PPR) exam, measured at the EPP-year level. All specifications include EPP fixed effects and year fixed effects. The reference year for the event study is 2024–25, the year immediately preceding academy implementation for the second cohort. The joint F-test reports the *p*-value from a test that all pre-period interaction coefficients are jointly zero. EPPs that adopted SOR Academies in 2024–25 are excluded from both groups. † Certification exam data through March 1, 2026. Standard errors clustered at the Educator Preparation Program level in parentheses.

$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Appendix B. Texas Teacher Certification Exam Descriptions

This appendix provides an overview of the Pedagogy and Professional Responsibilities (PPR) and Science of Reading (SOR) Texas teacher certification exams referenced in this study, including their domain structures and representative sample items.

Pedagogy and Professional Responsibilities EC–12 (Test 160)

The Pedagogy and Professional Responsibilities (PPR) EC–12 exam assesses candidates' knowledge of instructional design, classroom management, and professional conduct. The exam is organized into four domains:

1. **Domain I: Designing Instruction and Assessment to Promote Student Learning.** This domain addresses human developmental processes, learner diversity, instructional planning, and factors that impact student learning. Topics include developmentally appropriate practice, differentiated instruction, curriculum implementation, and critical thinking.

2. **Domain II: Creating a Positive, Productive Classroom Environment.** This domain covers establishing a safe and productive classroom climate, organizing the learning environment, and managing student behavior. Topics include physical classroom arrangement, fostering a passion for learning, formative monitoring, and cooperative group work.

3. **Domain III: Implementing Effective, Responsive Instruction and Assessment.** This domain focuses on communication strategies, active student engagement, technology integration, and monitoring student performance. Topics include questioning techniques, English-language learner supports, BYOD policies, and providing quality feedback.

4. **Domain IV: Fulfilling Professional Roles and Responsibilities.** This domain addresses family involvement, professional development, and legal and ethical requirements for Texas educators. Topics include parent partnerships, mentoring, vertical and horizontal teaming, mandatory reporting of suspected abuse, and IEP development.

Example items:

> *Which of the following activities would be the most developmentally appropriate for a first-grade teacher to include in a unit on the life cycle of plants?*
> (A) Planting a bean and observing the stages of growth
> (B) Reading nonfiction books about plants and creating posters
> (C) Visiting a science center and touring the plant exhibits
> (D) Watching a video about plants and drawing the bean life cycle

Which of the following instructional practices best ensures that a teacher structures the learning environment to address the needs of all students?
(A) Differentiation (B) Remediation (C) Enrichment (D) Modeling

Which of the following is the most appropriate first step to take if a teacher is concerned that a student is suffering from abuse?
(A) Calling the parents for a conference to discuss the suspicions
(B) Reporting the suspicions to the child protective services agency
(C) Sharing the concerns with the designated school personnel
(D) Asking the student pointed questions to ensure that all is well at home

Science of Teaching Reading (Test 293)

The Science of Teaching Reading (STR) exam assesses candidates' knowledge of evidence-based reading instruction for early childhood through grade six. The exam is organized into three domains:

1. **Domain I: Reading Pedagogy.** This domain covers foundational concepts in the science of reading, including assets-based approaches, dyslexia intervention, and reading assessment. Topics include systematic explicit instruction, formative and diagnostic assessment, and prevention of reading difficulties.

2. **Domain II: Reading Development—Foundational Skills.** This domain addresses phonological and phonemic awareness, print concepts and alphabet knowledge, phonics and word identification, syllabication and morphemic analysis, and reading fluency. Topics include phoneme-grapheme mapping, the alphabetic principle, decodable texts, inflectional endings, prosody, and automaticity.

3. **Domain III: Reading Development—Comprehension.** This domain covers vocabulary development (including Tier Two and Tier Three words), reading comprehension strategies, comprehension of literary texts, and comprehension of informational texts. Topics include background knowledge activation, text structure, metacognitive strategies, and scaffolding for English learners.

Example items:

A first-grade student has been identified as having dyslexia and has begun intervention. Which of the following approaches to instruction would be most effective to enhance the student's reading development?
(A) Allowing the student to use colored overlays on all classroom texts
(B) Using reading materials with specialized fonts designed for people with dyslexia
(C) Arranging for the student to use a working-memory training program daily
(D) Providing systematic, explicit multimodal instruction in all essential, evidence-based components of reading

Which of the following sets of words would be most appropriate to categorize as Tier Two words?
(A) fossil, air, soil (B) clock, book, floor
(C) arrange, observe, predict (D) custom, shelter, community

A third-grade English learner has grade-level decoding skills and scores around the grade-level benchmark for oral reading fluency. However, the student's text comprehension is mixed. When the student is having difficulty with a text, the teacher's best initial response should be to:
(A) Provide the student with a list of probing questions to answer after reading
(B) Talk with the student to informally assess background knowledge about the topic
(C) Advise the student to read more slowly and focus on comprehension
(D) Encourage the student to develop a written summary while reading

Score Reporting to Candidates

Candidates receive TExES score reports through their Pearson account beginning at 10:00pm Central time on the score report date, typically within 7–10 business days for selected-response exams; exams with constructed-response components take longer because trained scorers grade those responses. Figure B1 presents an example report for the English Language Arts and Reading 7–12 exam (Test 331). The report displays the candidate's total scaled score, pass/not-pass status, the scaled score range (100–300), and the minimum passing score (240). It also provides a performance breakdown by domain and competency, reporting points earned out of points possible for each, with constructed-response items reported separately. Under Texas Education Code, candidates are limited to five attempts per exam, but may appeal this with the Texas Education Agency.

Texas
Educator Certification
Examination Program

This barcode contains unique candidate information for security purposes.

Examinee Score Report

Test: 331 ENG LANGUAGE ARTS AND READING 7-12

Total Scaled Score: 262
Status: Passed
Scaled Score Range: 100-300
Passing Score: 240

Test Date: MM/DD/YYYY
FIRSTNAME M LASTNAME
123 SAMPLE STREET
CITY, TX 99999

TEA ID: 9999999

Performance by Domain and Competency	Points Possible	Points Earned
I. Reading Instruction and Assessment	23	14
001 Foundations of Reading Instruction and Assessment	8	4
002 Vocabulary Development	8	5
003 Reading Comprehension	7	5
II. Text Comprehension and Analysis	14	8
004 Reading Literary Texts	7	3
005 Reading Informational and Argumentative Texts	7	5
III. Oral and Written Communication	23	12
006 Composition	8	3
007 Inquiry and Research	8	4
008 Listening and Speaking	7	5
IV. Educating All Learners and Professional Practice	12	10
009 Differentiation Strategies in Planning and Practice	4	3
010 Culturally Responsive Practices	4	3
011 Data-Driven Practice and Formal/Informal Assessment	4	4
V. Constructed Response	8	8

Your results for the constructed-response item are reported above. For more information about the constructed-response item, see the preparation manual for this exam available on the testing program website at www.tx.nesinc.com.

See next page for additional information.

Page 1

Figure B1: Example TExES Score Report

Notes: This figure shows an example score report for the TExES English Language Arts and Reading 7–12 (Test 331) exam. The report displays the candidate's Total Scaled Score, pass/not-pass status, the scaled score range (100–300), and the minimum passing score (240), followed by a Performance by Domain and Competency breakdown reporting points earned out of points possible for each domain and competency. Source: Texas Educator Certification Examination Program (www.tx.nesinc.com).