Decoding Success: Evidence from a Federal Phonics Initiative*

Alex Hollingsworth, Alberto Ortega, and Nikolaos Zirogiannis†

September 26, 2026

Abstract

Learning to read is a foundational skill, yet fewer than one in three US fourth graders read at a proficient level. How best to teach reading and how to improve educational attainment has been the subject of an active policy debate for decades. Some argue for unrestricted funding increases, while others support more targeted changes to specific pedagogy. Recently, phonics-based approaches have surged in popularity, even though the largest attempt to put similar methods into classrooms was previously abandoned due to concerns that the program was ineffective. This paper studies Reading First—a nearly $1 billion per year federal program established under No Child Left Behind that promoted phonics-based instruction in K–3 classrooms. Using administrative data from Texas and exploiting the program's eligibility formula to construct a simulated instrument for district-level funding, we estimate causal effects via both a difference-in-differences design and a two-stage least squares framework. We find that Reading First eligibility increased third-grade reading test scores by 0.1 standard deviations. These effects persist through eighth grade, spill over to math scores, and are concentrated in high-poverty schools. District expenditure data show that Reading First dollars were spent on reading materials and teacher salaries. An analysis of digitized exam questions indicates that Reading First students outperformed peers on longer questions and those that required advanced decoding skills. Finally, we follow exposed cohorts into early adulthood and find that Reading First increased on-time high school graduation, four-year college completion, and adult earnings at age 25 by $554.

JEL Classifications: A20, H75, I22, I24

Keywords: Literacy, Phonics, Reading First, School Spending

*We would like to thank numerous conference participants at APPAM; the DC Health Economics seminar series; Jinan University; Indiana University, Instituto Tecnológico Autónomo de México; Miami University; North Carolina State University; Ohio State University; Saint Vincent College; the Southern Economic Association Annual Meetings; Syracuse University; University of Buffalo; the University of Kentucky; the University of Massachusetts - Amherst; the University of Pittsburgh; Vanderbilt University; Washington State University; and West Virginia University for their helpful feedback and suggestions. We also thank Jared Pine, Sungha Jo, Victor Xue for invaluable research assistance. Finally we thank Stéphane Lavertu and Vlad Kogan for their guidance and help throughout the project.

†Hollingsworth: Department of Agricultural, Environmental, and Development Economics & Department of Economics, Ohio State University and NBER (hollingsworth.126@osu.edu); Ortega: O'Neill School of Public and Environmental Affairs, Indiana University - Bloomington and NBER, (alorte@iu.edu), and Zirogiannis: O'Neill School of Public and Environmental Affairs, Indiana University - Bloomington, Email: nzirogia@iu.edu. The authors declare no known conflicts of interest, financial or otherwise, related to this work. We are grateful to the Spencer Foundation for supporting this research through Grant #202400025.

1 Introduction

Learning to read is a foundational skill. It is a cornerstone of early education around the world, as gains in early literacy can compound through later skill formation enabling later educational attainment, higher earnings, and greater economic mobility (Cunha et al., 2010; Chetty et al., 2011; Denning et al., 2023). Yet in 2024, fewer than one in three US fourth graders read at a proficient level, and the trend is worsening (NCES, 2024). The decline in reading proficiency amid rising educational expenditures highlights a critical need to understand how we allocate resources to generate both short- and long-run returns.

Society has many levers to improve literacy achievement. Within the classroom, research highlights the critical roles of teacher quality and class size, among other inputs (Hoxby, 1996; Krueger, 1999; Chetty et al., 2014b).[1] Moreover, additional school funding has been shown to raise achievement on average (Jackson et al., 2016; Jackson and Mackevicius, 2024), but there is less agreement over which specific investments produce gains, or whether districts do better with unrestricted funds (Abott et al., 2020; Baron, 2022). Despite this large literature, the specific pedagogical methods used in the classroom are notoriously difficult to observe in administrative data and thus have received relatively less attention. Curriculum and instructional materials have been characterized as the "forgotten" inputs in the education production function, frequently overshadowed by more expensive structural reforms (Whitehurst, 2009; Boser et al., 2015).

American classrooms have long been divided over how to teach reading. Phonics-based instruction teaches children to map letters to sounds, while "whole language" approaches emphasize context clues and text immersion (Goodman, 1967; Clay, 1991). The divide grew entrenched enough to become known as the "Reading Wars" (Mathews, 1966; Pearson, 1989; Kogan, 2025). Decades of research, ranging from laboratory experiments to brain imaging, find that children learn to read most effectively through phonics (National Reading Panel, 2000; Rayner, 1998; Castles et al., 2018). However, the largest previous attempt to put these methods into classrooms was abandoned due to concerns that the program was ineffective (Gamse et al., 2008). Nonetheless, a recent "science of reading" movement has surged in popularity, with over 40 states enacting legislation mandating phonics-based instruction in early literacy curricula (Schwartz, 2022; Kogan, 2025). States are thus rapidly scaling a pedagogy that is proven in the laboratory but of uncertain effect at scale.

What makes a reading intervention succeed at scale? How do pecuniary investments targeted at reading compare to unconditional school funding? Which students gain the most, and do their gains spill over into subjects beyond reading? Do the effects grow as instruction takes hold, or fade out like those of many early interventions? And do any of the gains persist long enough to affect labor market outcomes?

[1] Beyond the classroom, numerous studies examine external factors such as socioeconomic status, neighborhood violence, pollution exposure, nutrition, and the home language environment (Reber, 2010; Dahl and Lochner, 2012; Ang, 2021; Persico and Venator, 2021; Cabral et al., 2021; Hollingsworth et al., 2022; Cox et al., 2026; Ortega and Ludwig, 2022).

To answer these questions, we evaluate one of the largest government-led efforts to reshape classroom reading instruction in U.S. history– the Reading First program. Established under the No Child Left Behind Act, Reading First distributed nearly $1 billion annually nationwide to promote phonics-based instruction in Kindergarten through third-grade classrooms. We specifically examine the program's implementation in Texas, which received nearly $100 million per year. Beginning in late 2003, the state provided grants to eligible school districts, with funds designated for phonics-based curricula, new reading materials and textbooks, teacher coaching and professional development, diagnostic assessments, and additional staff.

To identify the causal effect of Reading First, we exploit the quasi-experimental variation created by funding rules in Texas. Program eligibility was determined by pre-existing test scores and district poverty indicators and the amount of support received by eligible districts was dictated by a funding allocation formula. We leverage this structure to drive two complementary identification strategies: a difference-in-differences approach that compares eligible and non-eligible districts before and after the program's start, and an instrumental variables framework that uses the funding formula as a simulated instrument for district-level funding. We execute this analysis using detailed administrative data from the Texas Education Research Center, tracking individual students from K-12 through postsecondary enrollment and into the labor market.

We find that Reading First eligibility increased third-grade reading test scores by approximately 0.1 standard deviations relative to non-eligible districts. The effects are evident immediately after the funding is first disbursed, grow during the funding period, and persist at a positive and statistically significant level even after the last Reading First dollar is spent. Moreover, students exposed to the program in early grades continue to relatively outperform their peers on the state's standardized reading exam through eighth grade. We also find positive effects on third-grade math scores, driven by questions with more words, suggesting that improvements in reading skills spill over to performance in other subjects. Event-study estimates show no pre-trend differences between eligible and non-eligible districts, and our results are robust to a battery of specification checks including alternative demographic controls, unbalanced panels, varying bandwidths around the eligibility cutoff, and clustering at various geographic levels. Our instrumental variables approach shows similar results, indicating that districts receiving more revenue per student experienced larger gains, with the marginal benefit of the treatment from additional dollars decreasing with spending.

Consistent with the program's design, which prioritized high-poverty areas, we find that effects are larger in schools with above-median shares of economically disadvantaged and Title I students.[2] The effects of Reading First are largest for schools with a larger share of Hispanic students. Eligible districts received roughly $75 per K–3 student per year over the post period (about $132 per student-year during the program's funding period), and the program produced

[2]Title I, Part A is a federal program under the Elementary and Secondary Education Act that provides financial assistance to Local Education Agencies and schools with high numbers or high percentages of children from low-income families.

a pooled improvement of approximately 0.1 standard deviations in reading test scores.

Even attributing the entire during-program cost of $132 per student-year to the smaller during-program effect of 0.07 standard deviations, the implied cost per 0.1 standard deviations is $188 per student-year, a fraction of the $3,000 per pupil per year that Jackson and Mackevicius (2024) estimate is typically required for a comparable gain. Under a marginal value of public funds framework of Hendren and Sprung-Keyser (2020), we find that Reading First pays for itself. The tax revenue recovered from students' higher adult earnings covers the program's full cost by age 30. Our setting also allows us to explore the effect of spending at both the extensive and intensive margins. We find that the marginal impact of each dollar declines with the amount received. Districts that received roughly eighteen times as much funding as the least-funded districts experienced treatment effects that were only about two and a half times as large.

We explore several channels through which Reading First may have improved achievement. First, we digitize and classify the content of third-grade reading and math exams and show that Reading First students outperform their non-treated peers on questions with more words, items testing vocabulary, questions requiring broad comprehension, and questions with a larger share of words that require advanced phonics decoding skills. The gains are concentrated in decodable content, and the same advantage appears on math questions that require reading and vanishes on those that do not. These findings are consistent with the hypothesis that phonics instruction improves decoding skills, which in turn help students engage with more demanding text. Second, using detailed district expenditure data, we show that Reading First funds were spent primarily on reading materials, general instructional supplies, and staff wages. We find little evidence that the funds changed the size or composition of the teaching workforce. Third, we find no evidence that schools in Reading First districts attracted more or higher-quality teachers, as measured by experience, education, or campus-level value-added. We then compare Reading First expenditures to a more general revenue shock, the passage of school district tax elections, and find that election-based discretionary funds flow primarily to salaries rather than reading materials and produce little improvement in test scores within our setting. This contrast suggests that Reading First's achievement effects reflect its curricular mandate rather than a simple infusion of additional money.

We follow the earliest Reading First-exposed cohorts as they age, first through high school and then into their mid-20s, and find that short-run achievement gains translate into lasting improvements in human capital. Eligible students are more likely to graduate from high school on time and more likely to enroll college, graduate from a four-year colleges, with no difference in two-year college enrollment or graduation. We also find that eligibility raises earnings at age 25 by $554, with most of that gain running through improved college enrollment and occupational upgrading.

Our work makes several contributions to the literature. First, we provide causal estimates of Reading First's effect on student outcomes. Despite being the academic cornerstone of No Child Left Behind and costing over $1 billion per year nationally, the program has received

little attention in the economics literature. Prior work commissioned by the federal government found no measurable effect on student reading comprehension (Gamse et al., 2008). While this null finding contributed to Reading First's defunding and shaped a decade of federal education policy, the evaluation was published in 2008 using only data prior to that year. Because Reading First was a K-3 intervention and the first funding was released in mid-2003, only a small fraction of the third graders tested in 2007 would have been "fully treated." Our study overcomes this limitation since we observe students who experienced the full effect of the program, examine the intensity of treatment, and consider longer run effects. This provides a more accurate and comprehensive assessment of Reading First's true efficacy.

Second, we present evidence on whether the scientific consensus on phonics instruction can be successfully implemented at scale through government policy. A growing body of work examines recent state-level reading reforms. Spencer (2024) finds that Mississippi's 2013 comprehensive literacy law substantially improved National Assessment of Educational Progress scores, gains dubbed the "Mississippi Miracle," though Wainer et al. (2026) argue that concurrent retention policies complicate causal attribution. Novicoff and Dee (2026) document positive effects of California's Early Literacy Support Block Grant, while Westall and Cummings (2023) find that only states adopting fully comprehensive literacy policies see reading gains. Descriptive evidence from Michigan's 2016 "Read by Grade 3" law also suggests phonics-focused reforms can coincide with score improvements (Strunk et al., 2021). Internationally, Machin et al. (2018) study the shift to synthetic phonics in England, and Álvarez Marinelli et al. (2024) evaluate a phonics-based tutoring program in Colombia. Our study complements this literature by examining the first major wave of government-mandated phonics instruction in the United States, in a setting that predates the recent "science of reading" movement. Because our analysis covers a period before phonics gained widespread popularity, our control districts are less likely to have adopted similar practices independently, yielding estimates that more cleanly capture the effect of the policy rather than a broader secular trend. Our setting also avoids the confounding effects of the COVID-19 pandemic (Davis et al., 2021; Bacher-Hicks et al., 2021; Dee, 2024), which complicates the interpretation of studies examining more recent reforms (Westall and Cummings, 2023).

Third, we contribute to the literature on the persistence of educational interventions and identify the mechanisms through which Reading First improved outcomes. A well-documented pattern is that early interventions produce test score gains that dissipate within a few years, yet some of these same programs still improve long-run outcomes, creating a puzzle about what mediates the long-run effects (Chetty et al., 2011; Jacob et al., 2010; Bailey et al., 2020; Cascio and Staiger, 2012). Our findings diverge from this pattern as Reading First induced reading gains that persist through eighth grade and beyond, suggesting that learning to read proficiently at a young age generates a "fork in the road" in human capital accumulation (Hoxby, 2021). This is consistent with models of dynamic skill formation in which early skills generate compounding returns to subsequent learning (Cunha and Heckman, 2007; Cunha et al., 2010). At least within the context of this intervention, we do our best to provide evidence on

the mechanisms that caused improved performance. We show that a substantial share of Reading First funds were spent buying instructional and reading materials. We corroborate this interpretation through an analysis of exam content. Reading First students outperform their peers specifically on more complex reading passages and on questions requiring more decodability and broad comprehension skills. Finally, we exploit the panel structure of our data to show that subsequent cohorts of third graders in Reading First districts benefit even after federal funding ends, indicating that Reading First induced lasting changes in school-level instructional practice, not just temporary gains for directly treated students. Taken together, these findings are suggestive of Reading First changing what happens in the classroom in ways that outlast the funding itself.

Fourth, our work speaks to the economics of federal intervention in a decentralized education system and the design of intergovernmental grants (Oates, 1999). The United States operates thousands of school districts with deep traditions of local financing and governance (Hoxby, 2000; Alesina et al., 2004), yet the federal government has increasingly used conditional grants to steer local practice. This design question has received considerable attention in other policy domains, particularly health care and welfare policy (Baicker and Staiger, 2005; Knight, 2002), with more limited evidence in education (Gordon, 2004; Cascio et al., 2010). Much of that evidence finds that targeted federal dollars are diverted or crowded out by local responses. We show that Reading First was a conditional grant that worked as designed. Reading First dollars reached eligible districts, were spent on the practices the program required rather than crowding out other revenue, and produced lasting achievement gains.

Fifth, we ask whether funding tied to specific practices produces different effects than an equivalent unconditional revenue increase. A growing consensus finds that school spending improves educational and economic outcomes (Jackson et al., 2016; Abott et al., 2020; Baron, 2022; Duque, 2026), but less is known about whether the composition of spending matters. We address this question by comparing the effects of Reading First expenditures with the effects of untargeted increases in district revenue generated by successful local tax elections. Using a difference-in-differences design to analyze school district tax referenda, we show that general revenue increases go primarily towards salaries and do not translate into purchases of reading materials or instructional supplies. This contrast suggests that conditional grants tied to evidence-based practices may be substantially more productive than equivalent unconditional revenue.

The remainder of the paper is organized as follows. Section 2 provides background on the Reading Wars and the Reading First program at both the national and Texas levels. Section 3 describes our data sources, including the administrative education records, district expenditure data, digitized exam questions, and school district tax elections. Section 4 outlines our empirical approach. Section 5 presents our main results on reading achievement along with internal validity checks, and campus and student-level heterogeneity. Section 6 explores mechanisms beginning with the question decoding analysis (where we also present the math results), followed by expenditure patterns, heterogeneity in effects by how districts allocated

their awards, school and teacher characteristics, and a comparison of Reading First expenditures to other exogenous sources of school funding. Section 7 presents long-run results, following exposed students through grades 4–8, high school, college, and into the labor market. Section 8 presents Marginal Value of Public Funds estimates for Reading First and section 9 concludes.

2 Background

2.1 Phonics and the Reading Wars

Debates about how best to teach children to read date back to at least the early nineteenth century (Mathews, 1966). While the various methods are not mutually exclusive, one persistent tension has been between methods that prioritize linking written text to the sounds the letters make versus methods that prioritize context (Goodman, 1967; Clay, 1991).[3]

Phonics is a form of teaching reading that is centered on connecting written letters to the sounds those letters make when spoken aloud. Learning to sound out printed text by inferring the sounds the letter pattern makes is known as decoding, which is a core phonics skill (Castles et al., 2018). For example, a student would learn the sounds that letters *c*, *a*, and *t* make when read aloud and then learn to string these sounds together to read the word *cat*, without relying on any additional information or context.

Many approaches to reading instruction incorporate some amount of phonics. However, curricula centered around phonics often follow a planned sequence that builds from teaching simple letter-sound patterns to more complicated ones in a systematic manner (National Reading Panel, 2000). For example, it is not until after students understand the *a* sound in *cat* that words like *coat* can be introduced, since in the latter case the long vowel letter pair *oa* changes the sound the *a* produces.[4]

By the late 1950s, phonics was a popular method of reading instruction in the United States in part because of the success of the 1955 book *Why Johnny Can't Read* and Cold War anxieties about American educational standards. However, phonics' foothold was soon replaced as methods that prioritized context became more prominent in the 1970s. The most popular of these was the whole language approach. A central tenet of this approach is that children can learn to read in a similar manner in which they learn to speak, through repeated exposure and immersion to the type of text people actually read rather than through the incremental letter-sound pairing skill formation emphasized in phonics. By 1989, the prominent reading scholar and Dean of the Graduate School of Education at UC-Berkeley, P. David Pearson wrote that he had never "witnessed anything like the rapid spread of the whole-language movement.

[3]See Kogan (2025) for an authoritative account of the reading wars and their political dimensions.

[4]The Center for the Improvement of Early Reading Achievement (CIERA) created a simple way to estimate how decodable words are, given the letter and sound patterns in each word (Hoffman et al., 2002). The scores range from 1 to 8. The word *cat* has a decodability score of 2 since it is a consonant-vowel-consonant pattern that does not end in an *r* or an *l*, *coat* with the long vowel letter pair has a score of 5, while the word *caught* would be the most difficult to decode at a score of 8. We use the CIERA decodability scorer to determine if exam questions require advanced phonics to read in Section 3.2.

Pick your metaphor—an epidemic, wildfire, manna from heaven—whole language has spread so rapidly throughout North America that it is a fact of life in literacy curriculum and research" (Pearson 1989 quoted in Kogan, 2025).

The debate took on a partisan dimension, with phonics increasingly identified with the Republican Party and whole language with Democrats. By the late 1990s, more than 65 percent of state legislators introducing phonics bills were Republicans (Paterson, 2000). In 1997, Congress instructed the National Institute of Child Health and Human Development to convene an expert panel to synthesize the evidence on reading instruction. The resulting National Reading Panel (2000) endorsed systematic and explicit phonics instruction, along with emphasis on phonemic awareness, fluency, vocabulary, and comprehension.

2.2 National Reading First Initiative

The Reading First program was established in 2002 under the No Child Left Behind Act (NCLB) of 2001. Authorized at $900 million in its first year, Reading First funding rose to about $1 billion annually by the mid-2000s (McCallion, 2005). Reading First aimed to ensure that all children learn to read proficiently by the end of third grade (U.S. Department of Education, 2002). The implementation was from Kindergarten to third-grade. The design of Reading First was heavily influenced by the emerging consensus on "scientifically based reading research", which highlighted phonics instruction and cited the findings from the National Reading Panel (2000). By statute, Reading First required schools to use curricula and teaching practices based on phonics. The program was unprecedented in scale and, at its peak, over 5,000 schools in over 1,800 districts participated, implementing comprehensive reading reforms with Reading First support (U.S. Department of Education, 2002, 2011).

Reading First emphasized phonemic awareness, phonics, and decoding as key skills needed to foster early literacy. It required that local programs provide explicit, systematic teaching in each of these areas. Federal requirements were that all instructional materials, lesson designs, and assessments should be grounded in scientifically proven approaches based on these components (U.S. Department of Education, 2002). Beyond curriculum and instruction, Reading First funds supported trainings for teachers for how to implement this new approach as well as the hiring of reading coaches trained in phonics-based instruction.

Reading First was administered as grants to states, after each state educational agency submitted a detailed plan. States then awarded subgrants to school districts based on need and were instructed to prioritize districts with high poverty and low reading performance (McCallion, 2005). In practice, the program reached nearly every state and served thousands of high-need elementary schools nationwide, although the exact way funds were administered to school districts differed by state. The majority of initial Reading First awards were made to states between the summer of 2003 (after the 2002-2003 exams were taken) and 2005, with the first fully treated set of students to go through the program, taking the third grade exams in 2007.

No Child Left Behind also mandated an independent national evaluation of Reading First's effectiveness, conducted by the US Department of Education. An initial impact study examining the effects of Reading First between 2004 and 2007 for 17 school districts across 13 states found that the program led to changes in teaching practices. Schools receiving Reading First funds spent more time teaching phonics-based approaches to reading and provided more teacher training and coaching. The report did not find evidence of statistically significant gains in students' reading comprehension by grade three (Gamse et al., 2008).

A separate federal implementation study that expanded the evaluation to 24 states similarly found that Reading First schools devoted more time to reading instruction and adopted the required instructional elements. The report found that nearly 100 percent of Reading First schools had a reading coach by the end of the program, compared to 57 percent of non-Reading First schools. However, this report also found limited evidence that Reading First schools improved reading achievement relative to non-Reading First schools (U.S. Department of Education, 2011). Given that Reading First was not randomly assigned to schools across the country, the report appropriately cites this as a caveat when interpreting its findings. Furthermore, these studies were conducted using data from 2004 to 2007, a time period where the vast majority of students within the study had yet to complete the entire Kindergarten through third-grade Reading First program.

The lack of statistically significant findings in Gamse et al. (2008) prompted questions about whether an intensive focus on phonics and other basic skills was enough to boost reading comprehension. These null findings coupled with concerns about the program's administrative efficiency (McCallion, 2005; Klein, 2008) ultimately led to the end of Reading First.[5] In 2008, Congress cut Reading First's funding by roughly 60% and then eliminated it entirely in fiscal year 2009 (U.S. Department of Education, 2011; Klein, 2008; Robelen, 2011). Although the end of federal appropriations occurred in 2009, previously awarded funds continued to flow to and be spent by school districts for several more years.

2.3 Reading First in Texas

Texas was awarded over $630 million from the federal government to implement Reading First. The money was disbursed from 2003 to 2009. Consistent with the national priorities of the Reading First program, Texas centered its efforts on implementing the types of phonics-based curricula recommended by the National Reading Panel (2000).

Texas developed materials and trained technical assistance providers to work with Reading First funded districts to develop implementation plans tailored to the core reading program. Within these districts, schools selected to receive Reading First funds would also receive training in screening, diagnostics, and progress monitoring to identify emerging reading difficulties,

[5]There were claims of insufficient transparency regarding states' requirements for receiving Reading First funding, and some questions about conflicts of interest among officials and contractors who helped implement the program in its early years (Klein, 2008).

as well as small-group supplemental intervention for students identified as at-risk (Hezel Associates, LLC, 2007).

Figure 1A presents a timeline for the roll-out of Reading First in Texas, along with relevant events at the Federal level. Texas submitted its Reading First application in April 2003 and received an award notification in May of that year. School districts within the state then competed for subgrants under strict eligibility criteria, prioritizing high-poverty schools with low reading performance. As indicated in Figure 1B, the first district awards were made in the summer of 2003, with the majority of initial grants taking effect in 2004. At its peak, the program operated across roughly 700 campuses statewide, with the state receiving approximately $100 million per year in federal funds.

Of every federal dollar received, the state of Texas had to distribute at least 80% to school districts. The remainder could be kept by the state to facilitate Reading First operations via program development (e.g., seminars), technical implementation assistance, and administrative overhead. By the end of Fiscal Year 2009, all Reading First funds had been distributed from the federal government to the state of Texas. Figure 1C shows that the state abided by this 80% limit as school districts received and spent about 83 percent of the federal money, slightly above the 80 percent floor.

The grey solid line in Figure 1D indicates that by 2005, nearly all (98.5 percent) school districts that were ever awarded a sub-grant had received at least some funding. Because K–3 spans four academic years, a student must enter kindergarten in a treated district and remain through third grade to be "fully treated." The dashed line in Figure 1D tracks the share of third graders in treated districts who were fully treated in this sense. The first fully treated cohort enters third grade in 2006 and took exams in 2007. The share of fully treated cohorts continues rising until 2008, when all third-grade classrooms in treated districts were "fully treated," since they had been exposed to Reading First curricular interventions for their entire K–3 period. This is because a district first funded in 2005 would not produce fully treated third graders until 2008. Figure 1E shows the geographic distribution of school districts that received Reading First funding, by Texas county. We discuss how Texas distributed the money to school districts and what this money was spent on in more detail in Section 3.3.

3 Data

3.1 Administrative Education Data

We obtain student-level education data from the Texas Education Research Center (ERC) at the University of Texas at Austin. The ERC links administrative records from multiple Texas state agencies, allowing us to follow students from public school through postsecondary education and into the labor market. Our primary outcomes are standardized math and reading test scores for grades 3–8, drawn from the Texas Assessment of Academic Skills (TAAS, 1995–2002), the Texas Assessment of Knowledge and Skills (TAKS, 2003–2011), and the State of

Texas Assessments of Academic Readiness (STAAR, 2012–present). We standardize scores to mean zero and standard deviation one among all test takers of a given test, within each state-test-grade-year cell. Thus our estimates are relative to the statewide score distribution of the exam actually taken that year, and year fixed effects absorb any statewide level shifts across exam regimes. The ERC data also contain rich student demographic information, including race and ethnicity, sex, English as a Second Language (ESL) status, gifted status, eligibility for free or reduced-price meals, and special education status.

Our main analysis is conducted at the campus-grade-year level, with a primary focus on third grade, the first grade in which standardized test scores are administered, and the last grade directly targeted by Reading First. We also examine effects on grades 4–8, since exposure to Reading First instruction in early grades may persist into later years. The main estimation sample is a balanced panel of campuses observed in every year over the 1995–2017 period; schools that open or close during the study window are excluded from this baseline, though our results are robust to using an unbalanced sample, and to extending the analysis period through 2019. Table B1 presents descriptive statistics for the full balanced-panel sample of 1,128 districts versus the sample of districts within five percentage points of the state's 2002 average fail rate of 13.6%.

Our baseline estimation sample includes only *first-time* test takers. Around 2003, Texas adopted a test-based grade-retention policy that substantially increased the number of students who repeated third grade and retook the reading exam (Zhong, 2026). Because this change coincided with the start of Reading First, a sample that included repeat takers could confound the program's effect with mechanical changes in who took the exam. The stable student identifiers in the ERC data let us flag the first time each student takes a given grade's exam, and we build all campus-grade-year means from those first-time takers. Section 4.4 addresses the selection concern this restriction could itself raise and presents results estimated on the full sample of all test takers. In addition to test scores, we examine student absences and disciplinary outcomes. We consider these outcomes given that early reading difficulties have been associated with classroom disengagement and behavioral problems (Algozzine et al., 2011; McIntosh et al., 2008; Morgan et al., 2008).

To study long-run outcomes, we follow those that were first enrolled in third-grade before 2014 into their mid-20s. We then explore whether short-run reading gains translate into longer-run improvements in high school completion, college enrollment and attendance, and early-career earnings. These results allow us to trace a specific literacy intervention from elementary school grades through early adulthood. College enrollment and graduation data come from the Texas Higher Education Coordinating Board (THECB), which tracks enrollment and degree completion at public, community, independent, career, and health-related institutions in Texas. Enrollment data in universities outside of Texas come from the National Student Clearinghouse. For labor market outcomes, we use quarterly earnings records from the Texas Workforce Commission (TWC), which covers employment and wages. However, our ability to study long-run outcomes is constrained by data availability. The first full-year students were

exposed to Reading First was 2004. Given that the labor data available to use, currently run through 2025, the oldest cohort that could have been exposed to Reading First in third grade — those in third grade in 2004, were only around 29 years old in 2025. The first cohort "fully" exposed to Reading First across all of K–3 (entering kindergarten in 2004) were approximately 25 years old in 2025. As a result, our labor market estimates capture early career outcomes rather than peak earnings, and should be interpreted accordingly.

Finally, the Texas Education Research Center data contain detailed information on school employees and wages, which we use to examine whether Reading First changed teacher composition and compensation across treated schools and districts. One important limitation is that we are unable to link students to teachers until 2012, after the Reading First program ended. We therefore construct campus-level measures of teacher quality to examine whether Reading First altered the distribution of teacher effectiveness across schools (Foy, 2024; Bacher-Hicks and Koedel, 2023; Biasi, 2021; Chetty et al., 2014a). We discuss this analysis further in Section 6.3.

3.2 Exam Questions

To more directly examine whether Reading First affected the linguistic demands of the assessments students faced—and to study heterogeneous effects by question type—we collected and digitized copies of the third-grade TAAS, TAKS and STAAR reading and math exams.

Since phonics instruction is tightly related to decoding sounds from letters, the composition of exam questions themselves may matter for how Reading First-exposed students perform. A student who has received phonics instruction may perform better on questions that contain words that require phonics skills to decode and on questions that necessitate close engagement with individual words (e.g., vocabulary questions). Conducting an analysis at the question-level enables us to test these hypotheses directly, going beyond average treatment effects and exploring if Reading First improved performance on exactly the type of questions where phonics skills are the most important.

We attempted to collect exams from all years in our sample from four sources: the Texas Education Agency website, the Internet Archive (*Wayback Machine*), the Texas Public Information Request process, and the Texas State Library. Figure A1 summarizes the years for which we obtain valid released 3rd grade exams and the number of questions by year. Exams from some years were available only as practice releases or could not be cross-validated against the ERC answer keys and are excluded from the analysis.

We digitize the full set of available 13 reading and 13 math exams and classify each question along four dimensions: 1) the total number of words in the question-and-answer pair, 2) whether the question asks students to define or identify the meaning of a specific word (a vocabulary item), 3) whether the question requires students to read an extended passage and demonstrate broad comprehension and 4) the share of words in a reading passage that required advanced phonics skills. We measure advanced phonics from the decodability of the words in each reading

question. Following the Center for the Improvement of Early Reading Achievement (CIERA) decodability scale, each word receives a score from 1 to 8. A score of 1 marks a word that can be sounded out with basic phonics, such as *me*, *we*, or *be*, while a score of 8 marks an irregular word that is very hard to decode and must be memorized (Hoffman et al., 2002). We collapse the scale into three groups. Basic phonics covers scores of 1 through 4, advanced phonics covers scores of 5 through 7, and irregular words receive a score of 8.

Our TEA data, record correct and incorrect responses at the student level, which we merge with the four question classifications discussed above to estimate heterogeneous treatment effects across item types. The four dimensions do not all apply to every question. Word count is defined for any reading or math exam question. The vocabulary, broad comprehension, and advanced phonics dimensions describe features of extended reading passages, so we code them only for reading questions.

3.3 District Revenues and Expenditures

We also obtain district-level revenue and expenditure data from the Texas Education Agency's Public Education Information Management System (PEIMS) Financial Data Reports, accessed through the TEA's publicly available financial data downloads (Texas Education Agency, 2024). The data include information on the fund code (source of funds), the function code (purpose of expenditure), and the object code (type of expenditure).

Reading First funds are tracked under a dedicated fund code, which allows us to separately identify both the amount of Reading First dollars received by each district and how those dollars were spent. We use the object codes to decompose Reading First expenditures into payroll, contracted services, supplies and materials, travel, and other categories, shedding light on the mechanisms through which Reading First may have affected student outcomes. Figure 1B plots total district Reading First fund expenditures by year, illustrating the rollout and wind-down of spending over the program period. The PEIMS also categorizes supplies and materials into books and general supplies. Books include reading materials, library books, and textbooks. The classifications within this budgetary data reflect the limited set of object codes contained within the dataset. For example, the category "General Supplies" contains purchases such as workbooks. While phonics-curricula promote the use of workbooks as a teaching material, we cannot accurately determine specific purchases of things like work books. We also cannot identify specific books (e.g., ISBN) purchased using these data. A final limitation of these data is that these are expenditures at the school district-fiscal year level. We do not have information at the school or grade level. Thus, we cannot specify how Reading First funds were spent across specific schools or employees within a school district.

3.4 School District Tax Elections

To examine how Reading First expenditures compare to the effects of broader budget shocks on district spending and outcomes, we leverage data on Texas school district tax elections

to raise discretionary operational funds from Abott et al. (2020), which cover 2000 to 2015, and extend the record through 2019 using districts' adopted tax rates (see Appendix C.1 for additional details). The passage of an election generates a plausibly exogenous increase in district revenues unrelated to any targeted programmatic objective. For this analysis, treated districts are those that passed a tax increase, and the comparison group consists of districts that had not yet passed one or never did.[6] It provides a plausible benchmark for assessing whether the effects of Reading First reflect something specific about how those dollars were spent (e.g., reading materials), rather than the infusion of additional resources alone. We merge the election data to PEIMS (discussed in Section 3.3) using district identifiers.

4 Empirical Approach

4.1 Predicting Eligibility

Given that receipt of Reading First funds may be non-random, we collect information on the components of the formula Texas used to determine which school districts were eligible for funding and how much money they would receive. We use those two formulas to predict district eligibility status and the amount of Reading First dollars a district received. Federal guidelines required that each school district receiving a competitive subgrant under Reading First must be allocated at least the same proportion of the state's total Reading First funds as it received under Title I, Part A in the previous year. Moreover, to ensure that all grantees received sufficient funding to successfully implement all activities in their Reading First plans, the state set a guaranteed minimum funding threshold of $170,000.

To qualify or become eligible, a school district had to meet two primary conditions. First, the district's Grade 3 failure rate on the 2002 standardized reading test had to be at or above the statewide average of 13.6 percent. Second, the district needed to oversee at least one of the following: (a) a geographic area designated as an empowerment zone or enterprise community under federal tax code; (b) at least one campus identified for school improvement under Title I Part A; or (c) a student population where either 15 percent or more of students, or more than 6,500 students, were eligible for basic Title I funding. We summarize Reading First eligibility (RFE_d) for district d as follows:

$$\begin{aligned} RFE_d = \mathbb{1}(\text{Fail Rate}_{d,2002} \geq 13.6\%) \times \max \Big[& \mathbb{1}(\text{LowPerforming}_{d,2002}), \\ & \mathbb{1}(\text{EmpowermentZone}_{d,2002}), \mathbb{1}(\text{Share of Title 1 Students}_{d,2002} \geq 15\%), \\ & \mathbb{1}(\#\text{of Title 1 Enrolled Students}_{d,2002} \geq 6,500) \Big] \end{aligned} \tag{1}$$

We can express the predicted Reading First dollars granted to a district d in year t as

[6] We do not use the close-election regression discontinuity design historically common in this literature for two reasons. First, there is recent evidence that conditioning on close races can bundle passage with the district characteristics that keep those races close (Marshall, 2024). Second, a difference-in-differences specification makes the tax-election estimates directly comparable to our main Reading First specification described in the following section.

$$\text{Grant}_{dt} = \max \left(\$170{,}000, \;\; \text{Reading First State Funds}_t \times \frac{\text{Title 1 Funds}_{dt}}{\sum_{j=1}^{N} \text{Title 1 Funds}_{jt}} \right) \times RFE_d \quad (2)$$

where N is the number of Texas districts with a Title I Part A allocation in the base year (1,206), and Reading First State Funds is the portion of Texas's federal Reading First appropriation in year t that was allocated to districts (80 percent of the total). Eligible districts get the maximum of $170,000 or a share of Reading First funds in year t equivalent to their share of Title I allocation. We use cumulative predicted Reading First dollars in levels rather than per student. The reason is mechanical. Scaling by enrollment places district size in the denominator of both the endogenous variable and the instrument. A small number of small districts with unusually high or low funding per student then generate variation that the eligibility formula does not predict. We prefer the unscaled measure because Reading First grants were awarded and spent in dollars, so it captures program intensity at the margin of an additional dollar. One limitation of our eligibility measure is that we cannot obtain data on all components of the eligibility formula for every year. Thus, we use 2004 as the start of the Reading First program and the eligibility criteria from 2002 to estimate the Reading First dollars for each year.

Figure 1F plots the first stage of our analysis, showing the relationship between cumulative predicted Reading First dollars allotted to a district and a district's cumulative Reading First expenditures. There is a clear positive relationship between the two variables.

4.2 Difference-in-Differences

To examine the effects of Reading First on the extensive margin, we employ a canonical difference-in-differences (DiD) two-way fixed-effects and event study model, where we use our estimated eligibility, from equation (1), as the treatment variable and the year the program effectively starts, 2004, as the pre-and post-demarcation year. Our identification strategy relies on the evolution of test scores or related outcomes before the introduction of Reading First. Since we cannot observe the counterfactual, inferring causality depends on the assumption that outcomes evolve similarly in Reading First-eligible and non-Reading First-eligible districts before Texas introduces Reading First. Thus, it would be reasonable to assume that a common trend pre-Reading First would have persisted in the absence of the program. Put another way, the timing of Reading First must be exogenous to pre-existing trends in our outcomes (e.g., test scores), and we should expect similar trends to exist in the absence of Reading First. If this holds, campus or district fixed effects will account for key time-invariant cross-sectional differences, and non-Reading First districts will capture trends in our outcomes over time, providing a counterfactual for how we expect our outcomes to evolve in the absence of Reading

First.

Suggestive evidence for the validity of our difference-in-differences identification strategy appears in Figure 2 panel A.[7] Here, we use the raw test score data to plot the difference in standardized third-grade reading test scores relative to 2003, between districts that, based on our prediction, were Reading First eligible vs ineligible. We see that reading scores in Reading First eligible districts are increasing over time relative to those in ineligible districts. The divergence is suggestive of the beneficial effects of Reading First on reading. Note that Figure 2A only exploits variation in location, not in timing. Thus, we employ the following difference-in-differences (DiD) event-study framework:

$$y_{cg(d)t} = \alpha_{cg} + \gamma_t + \sum_{\substack{j=1995 \\ j\neq 2003}}^{2017} \beta_j RFE_d \times \mathbb{1}\{t = j\} + \sum_{t\neq 2002} \psi_t \left(X_{cg}^{2002} \times \mathbb{1}\{\text{Year} = t\}\right) + v_{cg(d)t}, \quad (3)$$

where $y_{cg(d)t}$ is the mean standardized reading test score in campus c in grade g of school district d in year t. The terms α_{cg}, γ_t represent campus-by-grade and year fixed effects, respectively. RFE_d is an indicator variable equal to one if a district is predicted to be eligible for Reading First. The indicator variable $\mathbb{1}\{t - 2004 = j\}$ is equal to one if the observation year is j years from the start of Reading First in 2004. Event time j therefore runs from calendar year 1995 to calendar year 2017, with $j = 2003$ as the omitted reference period. Our main specification also controls for baseline campus composition by interacting the campus share of economically disadvantaged students in 2002, X_c^{2002}, with year fixed effects. Because Reading First targeted poor districts, these interactions allow test scores to trend differently by baseline poverty, so we compare eligible and ineligible campuses with similar shares of disadvantaged students. In the Appendix, we show that our results are similar, if not larger, if we exclude this control. We also find similar estimates if we include additional baseline campus demographic controls or control for time-varying campus demographics. For all specifications, confidence intervals are constructed from robust standard errors, clustered at the school district level.

Although Texas received its award notification in May 2003, we treat this year as a pre-treatment reference year. The first district sub-grants were not made until that summer, and because Texas administers its third-grade exams in the spring, the 2003 test was taken before any district had received Reading First funds. Panel B of Figure 1 confirms that district receipts begin only in fiscal year 2004.

The coefficients of interest in equation (3) are β_j, which capture the relationship between Reading First eligibility and test scores over time, both before and after a Reading First subgrant. An assumption of the difference-in-differences model is "parallel trends," in which the trends in the outcomes, $y_{cg(d)t}$, are common before Reading First. This test is embedded in the event-study framework, equation (3), where the difference in pre-policy trends is captured by β_j for $j \leq 2003$. The event-study design supports the common trends assumption if these

[7] Figure A2 shows the raw trend in standardized test scores without setting 2003 as a relative year.

pre-period coefficients are statistically insignificant and close to zero. The coefficients, β_j for $j \geq 2004$, allow us to examine the dynamic effects for the years after Reading First is introduced.

We summarize the average treatment effect using the following specification

$$y_{cg(d)t} = \alpha_{cg} + \gamma_t + \beta RFE_d \times \mathbb{1}(Year \geq 2004) + \sum_{t \neq 2002} \psi_t \left(X_{cg}^{2002} \times \mathbb{1}\{\text{Year} = t\} \right) + v_{cg(d)t}, \quad (4)$$

where $\mathbb{1}(Year \geq 2004)$ is an indicator equal to one for years after 2003. Given that Reading First spending ended in 2011, we also estimate a version of equation (4) that allows the treatment effect to differ across the two periods, replacing $\beta RFE_d \times \mathbb{1}(Year \geq 2004)$ with $\beta RFE_d \times \mathbb{1}\{2004 \leq t \leq 2011\} + \eta RFE_d \times \mathbb{1}\{t > 2011\}$, where β captures the effect during the program years and η the effect in the years following its conclusion.

We amend equations (3) and (4) to fit other outcomes accordingly. For instance, when examining absences or disciplinary outcomes $y_{cg(d)t}$ becomes the number of incidents per student in a campus-grade-year (where we use the campus-grade enrollment in 2002 as the denominator). When examining district-level revenue or expenditures, we estimate our regressions at the district-year level.

For long-run outcomes, we move from campus-year estimates to following exposed cohorts into high school, college, and the labor market (Section 7). We adapt the specifications above to the student-cohort level, defining treatment by the district a student attended in early grades and the cohort's exposure to Reading First. For test score, high school, and college outcomes, we estimate fixed-effects OLS models analogous to equation (4). For labor market outcomes, we estimate fixed-effects Poisson models which model zeros explicitly (Chen and Roth, 2024).

4.3 Two-Stage Least Squares

Given that we can observe Reading First dollar amounts and predict Reading First funds based on eligibility, we can also speak to the effect of Reading First at the intensive margin. Thus, we can use equation (2) as an instrument for the actual Reading First dollars a district received from the state. Given that school districts did not spend Reading First dollars every year, we examine the cumulative total of Reading First funds. For instance, a district may receive Reading First dollars in 2005 and spend them in 2006 and 2008, but not in 2007. Therefore, our first stage regresses the cumulative total of actual Reading First expenditures in a district through year t on the predicted cumulative amount. The resulting second stage can be expressed as follows

$$y_{cg(d)t} = \alpha_{cg} + \gamma_t + \theta \left(\sum_{s=2004}^{t} \widehat{RFE}_{ds} \right) + \sum_{t \neq 2002} \psi_t \left(X_{cg}^{2002} \times \mathbb{1}\{\text{Year} = t\} \right) + v_{cg(d)t}, \quad (5)$$

Where $\sum_{s=2004}^{t} \widehat{RFE}_{ds}$ captures predicted cumulative Reading First revenues obtained from the first stage. θ captures the total accumulated effect of one more dollar added anywhere be-

tween 2004 and year t. For ease of interpretation, when presenting the estimated effects on test scores, we scale θ by \$20 million (2025 dollars), roughly the cumulative Reading First dollars received by the average eligible student's district over the program (\$18 million). All other variables are as specified in Section 4.2. Our instrument is a simulated variable constructed from the eligibility formula, in the simulated-instrument tradition of Currie and Gruber (1996). As we show below, the first-stage F statistic far exceeds the conventional weak-instrument threshold in every specification we estimate (Appendix Table B4). With first-stage F statistics of this size, the weak-instrument-robust tF procedure of Lee et al. (2022) coincides with conventional inference.

The exclusion restriction is that, conditional on the fixed effects and controls, predicted Reading First dollars affect outcomes only through the Reading First dollars districts actually received. Level differences in size and poverty are absorbed by the campus fixed effects, and differential trends by poverty are absorbed by the 2002 economically disadvantaged share interacted with year fixed effects. The remaining concern, that predicted dollars track other funding after 2003, is tested directly in the Appendix, where we find no evidence that Reading First predicts non-Reading First revenue (Appendix Table B2) and that our main estimates are robust to controlling for non-Reading First district revenue sources (Appendix Table B6).

We also note that Reading First expenditures are observed at the district-year level. Thus, in our specifications, every campus in a district is assigned the same per-student expenditure. A district that concentrated its Reading First spending on a subset of campuses is treated identically to one that spread funds evenly. We do, however, examine whether higher-poverty campuses within a district gain more than their lower-poverty counterparts, as this was one of Reading First's stated goals. An alternative identification strategy would exploit the 13.6 percent failure rate cutoff sharply, comparing districts just above the threshold to those just below in a regression discontinuity design. We do not use this approach for several reasons. First, the eligibility criteria are multidimensional and incorporate more than just the 13.6 percent failure rate threshold. Thus, the fail rate is necessary but not sufficient for eligibility, as districts must also satisfy at least one of the secondary criteria in equation (1). Second, following all districts over time and comparing their trajectories before and after Reading First uses a broader and more representative set of comparisons, and allows us to compare campuses over time. Moreover, our 2SLS strategy allows us to speak to the effects of Reading First funding at the intensive margin, which varies continuously across eligible districts through the Title I allocation formula. Nonetheless, we present specifications where the estimation is restricted to districts near the 13.6 percent threshold, and find qualitatively similar results.

Our two main estimators complement each other but likely measure different aspects of Reading First. The difference-in-differences captures the extensive margin, the effect of becoming eligible, while the 2SLS captures the intensive margin, the effect of an additional dollar. If eligibility had mattered only through the dollars it brought, and every dollar had the same effect, the two sets of estimates would perfectly coincide. However, there are reasons to expect the effects at each of these margins to differ and to even expect non-linear effects due to

additional dollars. Part of the treatment arrived with eligibility rather than with dollars, as Reading First included a change in curriculum that is independent of the size of the grant a school district received. Moreover, there may also be diminishing returns to an additional dollar a district receives. Lastly, the two estimators weight districts differently, since variation in dollars is concentrated among the largest grants. We return to this distinction when presenting the results.

4.4 Threats to Identification

Our design requires that eligibility not coincide with other changes affecting reading outcomes. The clearest such concern is the rest of No Child Left Behind itself, since the characteristics determining Reading First eligibility may also govern districts' exposure to adequate yearly progress requirements, mandatory school choice, and supplemental educational services. If these accountability pressures improved reading instruction on their own, our estimates would bundle their effect with Reading First's. Our analysis limits this concern in a few ways. First, many of these provisions applied statewide. All Title I schools faced sanctions based on their own performance, in eligible and ineligible districts alike, and Reading First eligibility did not itself trigger any sanction. Second, our narrow-bandwidth specifications compare districts near the 13.6 percent fail rate cutoff component of the eligibility formula. Third, our heterogeneity analysis compares above-median Title I campuses on both sides of the eligibility boundary, which again holds the accountability environment roughly constant. We also consider a placebo test for the effect of eligibility on revenue outside of Reading First (Appendix Table B2), and estimate specifications that control for other revenue and for the 2002 socioeconomic composition of the campus interacted with year fixed effects.

A second concern is Texas's Student Success Initiative. Beginning with the 2002–03 third-grade cohort, this initiative required students to pass the state reading test within three attempts to advance to fourth grade. Zhong (2026) shows that retention under this policy temporarily raises retained students' subsequent scores while reducing high school graduation and adult earnings. Given that the Student Success Initiative was introduced one year before Reading First outcomes begin, we verify that none of its margins (money, targeting, or retention) drives our estimates. Because this initiative mandates retakes for students who fail, we score each student's first attempt throughout. In the appendix, we find similar effects when including all tests taken in a given year (Appendix Table B3).

5 Test Scores

5.1 First-Stage and District Revenues

We begin by establishing that Reading First Eligibility did indeed increase Reading First funding. Table 1 presents the relationship between Reading First eligibility and Reading First-specific revenue. Panel A presents our difference-in-differences estimates for the entire post-

period. Panel B splits the post-Reading First period into separate joint treatment effect periods for the years during the Reading First program (2004-2011) and for the years once the funding ended post-2011, respectively. Panel C uses our predicted Reading First dollars from equation (2) to estimate Reading First funding. Column (1) in Table 1 considers all school districts, while Column (2) restricts the analysis to school districts within five percentage points of the 13.6 reading test score fail rate (in 2002) used in the eligibility formula in equation (1). Across all specifications, we see that Reading First eligibility was predictive of district Reading First revenues received. In Column (1) Panel A, Reading First eligible districts received an additional $75 per K-3 student post-2003 relative to non eligible districts. Column (2) Panel A, suggests that the estimated effect is similar ($59 per K-3 student) when restricting the analysis to districts within five-percentage points of the 13.6 fail rate. Reassuringly, in Panel B we see that the estimated increase in Reading First revenues is confined to the period the program funding was operational. During 2004-2011, the period when Reading First funds were disbursed, Reading First eligible districts received $132 per K-3 student. Panel C (column 1) shows a strong first stage result for our instrumental variables approach with each additional $1 million in cumulative predicted Reading First dollars associated with $1.41 million increase in cumulative Reading First revenue, and a first-stage F statistic of 1,688.

5.2 Third-Grade

Figure 2B presents the pre- and post-treatment effects from equation (3), where the outcome is standardized third-grade reading test scores. The figure confirms the absence of pre-trend differences in third-grade reading between Reading First-eligible and non-eligible districts. After the start of Reading First in 2004, we see a clear trend break and a positive effect on reading test scores. The estimated effect increases until 2012, one year after the last Reading First dollars are spent. After this, the effect levels out and remains positive and statistically significant. Figure 2C re-estimates our main event-study specification but restricts the analysis to school districts within five percentage points of the 2002 13.6 percent fail rate used to determine eligibility for Reading First. This specification allows us to compare districts that are plausibly more similar at baseline. The bandwidth model shows a similar pattern.

Table 2 presents the average treatment effect on Reading First eligibility on third-grade reading scores. Column 1 of Panel A reports the pooled difference-in-differences estimate, indicating that campuses in eligible districts improved by 0.11 standard deviations after the program began. Panel B separates the treatment effect into the active funding and post-funding periods. The estimated effect is 0.07 standard deviations during 2004 through 2011 and 0.16 standard deviations after 2011. The large post-period effect is consistent with the event study estimates discussed above, which suggest that the effects of Reading First increased for several years and then leveled off. This pattern is consistent with the fact that there was an increasing share of each cohort who was treated across time. Panel D of Figure 1 shows the share of each cohort that was "fully" treated across time (meaning they had exposure to reading first in each

of Kindergarten through third-grade). It is not until 2008 that all treated students are "fully" treated in this sense.

Panel C of Table 2 examines the intensive margin by estimating the effect of cumulative Reading First dollars on third-grade reading. Column (1) reports the 2SLS estimate for the full sample, instrumenting actual cumulative Reading First expenditures with predicted Reading First dollars. The first-stage F-statistic far exceeds the conventional weak-instrument threshold (F = 1,711), and Appendix Table B4 shows that the first stage is similarly strong across different robustness specifications (F statistic between roughly 1,200 and 2,700 in the full sample, and above 750 in the narrow-bandwidth sample). As with our difference-in-differences model, we see that additional cumulative Reading First expenditures improve third-grade reading.

Our 2SLS estimates are smaller than our difference-in-differences estimates. At the average eligible district's cumulative funding ($20 million), the 2SLS reading effect is about half as large. The gap reflects what each estimate measures. The difference-in-differences captures the effect of eligibility for the average treated student, while the 2SLS captures the effect of the average dollar. Appendix Figure A17 estimates both models separately for eligible districts in three groups of cumulative predicted funding (under $5 million, $5 to $15 million, and $15 million or more). Districts in the high-spending group received almost twenty times the funding as those in the low-spending group. And while reading gains were higher for the high-spending group, they were only about two and a half times larger, not twenty times. Panel A shows the treatment effect by group at each group's average funding level for a treated student within that group. While Panel B instead rescales each estimate to be at the same dollar level. On a per-dollar basis, reading gains decline monotonically from the low-spending to the high-spending group. Within each spending group, however, the difference-in-differences and 2SLS estimates are very close to one another. The pooled estimates differ because the local average treatment effect is dominated by the high-spending group. This group contains the most students and the highest average spending. Here, each additional dollar produced the smallest gains. The low-spending group's per-dollar estimates are more imprecise and have a weaker first stage as a larger share of eligible districts that were never awarded actual Reading First dollars are contained within this group.

Beyond raising average scores, we also consider whether Reading First shifted students across the achievement distribution. Figure A3 plots the effect of Reading First eligibility on the probability that a first-time third-grade test taker places in each decile of the statewide reading distribution. Eligibility reduces membership in the bottom deciles - the probability of placing in the lowest decile falls by roughly three percentage points - and increases membership in deciles seven through nine by one to three percentage points. The pattern is similar when restricting to districts within five percentage points of the eligibility threshold, as well as when we consider shifts in the reading decile distribution of eighth graders. This distributional evidence indicates that the program moved struggling readers out of the bottom of the distribution rather than merely boosting students who were already proficient, consistent with the program's design of targeting struggling readers. Figure A3 also examines the probability of transitioning across

reading test score deciles between 3rd and 8th grade. For example, the very top left-hand corner cell suggests that conditional on being in the bottom decile of a reading test score, a Reading First third grader was 1.1 percentage points less likely to remain in that lowest decile by 8th grade, relative to a non-Reading First student. Overall, Panel C of Figure A3 provides evidence suggesting that low-performing 3rd-grade students had increased probabilities of moving to higher test score deciles by 8th grade.

We also examine the effect of Reading First on math in Figure A4 and find improvements in math test scores. In standardized terms, using our main equation (4), the math effect is 0.11 standard deviations, almost identical to the reading effect (see Appendix Table B5). This result is consistent with recent evidence suggesting that targeted interventions in reading (or math) may have spillover effects on math (or reading) (Xu et al., 2023). In Section 6.1 we investigate the potential mechanism that leads to this result. Specifically, we assess the plausible connection between improvement in reading and math competency by examining whether student success differs by the number of words in standardized math test questions. Appendix Figure A15 repeats the decile analysis for math and shows a similar pattern, with students moving out of the bottom decile and into the upper-middle deciles.

Finally, we examine three behavioral outcomes that might be influenced by improved reading performance, namely disciplinary actions, suspensions and attendance. As indicated by the event studies in the three panels of Figure A5 we do not find an effect of Reading First on those outcomes.

5.3 Internal Validity Checks

Panels D and E of Figure 2 present internal validity checks for our difference-in-differences and 2SLS models, respectively. Every row of each panel is a separate specification with the first row representing our baseline estimate. Row 2 removes the economically disadvantaged control from the baseline specification, leaving no controls in the regression. Row 3 adds 2002 demographic controls (percent Black, percent Hispanic, percent Asian/Native) interacted with year fixed effects. Row 4 replaces those interacted controls with time-varying demographics. Row 5 uses the same demographic controls as row 3 and adds the share of students who are gifted, or ESL, or who receive special education. Row 6 removes the balanced campus restriction and includes all campuses, resulting in an unbalanced panel of schools. Row 7 removes any student weights. Texas school districts lie within 20 Education Service Center regions which provide services to improve school performance.[8] Given that Reading First districts interact with these regional centers and that this may spill over to other districts, row 8 clusters standard errors at the region level (instead of district) and row 9 includes region-by-year fixed effects. During our sample period Texas changed their exams twice.[9] Row 10 excludes these exam

[8]Texas created Regional Education Service Centers to help school districts improve student outcomes, operate more efficiently, and carry out state-mandated initiatives. They support local capacity while serving as intermediaries between districts and the Texas Education Agency.

[9]Texas replaced the Texas Assessment of Academic Skills (TAAS) in 2003 with the Texas Assessment of Knowledge and Skills (TAKS) and then replaced TAKS with State of Texas Assessments of Academic Readiness

transition years. Row 11 keeps years before 2012, when Reading First funding was available. The last row includes all test takers, not just students who take the reading exam for the first time. Appendix Figure A16 plots the full event study for each of these specifications and reports positive effects on third-grade reading similar to those of the baseline specification.

Given that the probability of becoming eligible for Reading First jumps discretely at the 2002 third-grade reading fail rate of 13.6, we can compare districts within this bandwidth to improve the comparison between treatment and control. Figure A6 plots the average treatment effect estimate from our difference-in-differences and 2SLS specifications for different bandwidths around the 13.6 percent fail rate cut-off. Overall, we see that our results are positive irrespective of bandwidth. The effects increase with the bandwidth up to about twelve percentage points and then level off, with the difference-in-differences estimates less precise for the narrowest bandwidths of three percentage points and below. The first-stage relationship is also stable across every robustness specification, with instrument F-statistics well above conventional weak-instrument thresholds (Appendix Table B4).

We also contend with the possibility that districts received other revenue at the same time that Reading First funds arrived. Reading First was one program within the broader No Child Left Behind Act, and the district characteristics that determined eligibility may have also shaped exposure to other government funding streams. If eligible districts were awarded other grants or general revenue in the same years, our estimates could in part reflect the impact of those resources rather than Reading First itself. We address this concern in two ways. First, we test whether Reading First led to changes in non-Reading First revenues. Although Reading First may correlate with district factors or lead to changes that can subsequently affect a school district's revenue, finding a small or no effect on non-Reading First revenue lends additional credibility to our identification strategy and to improved reading instruction as the operating channel driving our findings. We present the effects of Reading First on non-Reading First revenue in Appendix Table B2. We find no economically or statistically meaningful effect of Reading First on non-Reading First district revenue. Second, we re-estimate our main specification while controlling directly for non-Reading First revenues and find that our estimates are unaffected (Appendix Table B6).

Lastly, we address the possible threat of selection-through-retention. Starting in 2002–03, Texas required third graders to pass the state reading exam to advance, retaining students who failed three attempts. Zhong (2026) studies this policy directly and finds that retention temporarily raises the test scores of retained students, so average scores can rise mechanically as the mix of test takers changes. Our main specifications address this by restricting the sample to first-time test takers, and we verify that Reading First itself did not change retention behavior (see Figure A7). The first-attempt restriction removes any retake-driven change in composition. As a final check, we re-estimate our main results on the sample of all test takers in Table B3 and Figure A8 and find effects very similar to our baseline specification.

The Student Success Initiative also carried dedicated state funding, so a natural concern

(STAAR) in 2012.

is that Reading First eligibility coincided with a differential flow of these dollars. We find no evidence of this on either margin. Eligibility predicts neither whether a district received any funds from this initiative nor how much it received. The dollar estimate is statistically insignificant and negative near the 13.6 percent fail rate threshold, and the event studies show no differential movement in Student Success Initiative funding as a result of Reading First's rollout (see Appendix Table B7 and Appendix Figure A9).

5.4 Heterogeneity

Figure 3 summarizes heterogeneity in the Reading First effect along two dimensions. Panels A and B examine variation across campuses based on 2002 campus characteristics, and Panels C and D examine variation across students within campuses based on individual characteristics. For each dimension, we report both difference-in-differences estimates and the corresponding IV estimates per $20 million of Reading First dollars (2025 dollars), for the full sample and for districts within five percentage points of the eligibility threshold.

Beyond documenting heterogeneity in treatment effects, the exercise in Panels A and B of Figure 3 has the added benefit of comparing campuses within Reading First and non-Reading First districts that are more similar to one another along observable dimensions. Given that a central goal of Reading First was to improve reading outcomes in high-poverty schools, we are particularly interested in whether treatment effects are larger for campuses serving more disadvantaged populations. The program's funding formula itself directed resources through the Title I allocation, so we would expect effects to be concentrated among higher-poverty campuses if the program operated as intended. We begin with the extensive-margin estimates in Panel A. Rows two through five examine heterogeneity by racial and ethnic composition. Campuses above and below the median share of Black students show similar point estimates, as do campuses split by Asian-Native enrollment. The clearest difference by race or ethnicity appears for the Hispanic share (the largest minority group in Texas during our sample period), where campuses above the median show a notably larger effect than those below it. The next row corroborates this pattern from the other direction, as campuses with above-median white enrollment show a smaller effect. Rows six and seven turn to measures of economic disadvantage. The Reading First effect is larger for campuses with above-median Title I shares and above-median shares of economically disadvantaged students, while the corresponding below-median estimates are smaller. Thus, at the extensive margin, we find that Reading First had its largest effects in higher-poverty schools with larger shares of disadvantaged students. These are the students the funding formula most directly targeted, so the program appears to have reached the population it was designed to serve. In Panel A we also find similar results when stratifying by the share of students classified as special education, gifted, or ESL. The intensive-margin estimates in Panel B are more uniform. Per-dollar improvements are similar for campuses above and below the median Title I share, so the concentration of effects among higher-poverty campuses is a feature of eligibility rather than of the marginal dollar.

Panels C and D stratify by student characteristics. Point estimates are positive for every racial and ethnic group. In the difference-in-differences specification, the gains are broadly similar across racial groups. In the 2SLS specification, the per-dollar effects are largest for Hispanic students and relatively smaller for Black and White students. We also consider a specification where the outcome is the difference between Black-white and Hispanic-white test scores, respectively. Consistent with these patterns, the gap outcomes at the bottom of Panels C and D show a narrowing White–Black gap in the difference-in-differences specification and a narrowing White–Hispanic gap in the 2SLS specification, which provides suggestive evidence that Reading First funds reached the disadvantaged populations it targeted. Positive effects of Reading First on reading are similar for both male and female students. Effects for ESL students are imprecise. Consistent with the campus-level heterogeneity analysis, economically disadvantaged and Title I students show relatively larger test-score gains. Special education students also show large gains, which fits a program that reaches struggling readers, while gifted students gain less.

Lastly, motivated by a growing literature documenting the effects of air pollution on both short-run and long-run student performance, Table B8 also splits campuses by median $PM_{2.5}$ concentration (Currie et al., 2009; Ebenstein et al., 2016; Hollingsworth et al., 2022; Persico and Venator, 2021; Jacqz, 2022; Yoder, 2024). The difference-in-differences estimates are similar across the two groups. The 2SLS estimate is larger below the median, but it rests on a much weaker first stage. We note that pollution exposure is correlated with economic disadvantage and demographic composition in this setting.

6 Possible Mechanisms

6.1 Decoding Test Questions

As described in Section 3.2, we digitized third-grade reading and math exams and classified each question by word count, whether it tests vocabulary or broad comprehension, and whether it involves advanced phonics. We examine the effects of Reading First on the share of correct responses on third-grade exams at the question-campus-year level as the outcome. Table 3 presents results from our main difference-in-differences specification. Columns (1)-(2) pool math and reading exams together, columns (3)-(4) report results for math and columns (5)-(10) results for reading exams. Panel A of Table 3 reproduces our main finding (illustrated in Table 2) using the share of questions answered correctly rather than standardized scores. We take the average share of correct answers in every question of a third-grade reading or math exam by campus-year which makes our unit of observation the campus-grade-year-exam question. Consistent with our earlier results, Reading First students outperform non-Reading First students in both subjects.

In Panel B, we interact the Reading First treatment with the word count of each question. The negative coefficient on word count indicates that all students perform worse on longer

questions. However, the positive Reading First interaction shows that this penalty is relatively smaller for students in Reading First-eligible districts. For each additional word in a question, Reading First-eligible students scored 0.03 percentage points higher than non-Reading First-eligible students (column 1). Panel C examines reading comprehension in more detail. Column 5 focuses on vocabulary questions, which test whether students know the meaning of specific words in a reading passage. Our results show that, while all students do better in vocabulary-related questions, Reading First-eligible students score 2.17 percentage points higher than non-Reading First-eligible students. Column 7 examines if Reading First students do better on questions that require broad comprehension. These items are among the most challenging on the exam since they test students on their overall understanding of the main message from an entire 2-3 page reading passage. Students score roughly four percentage points lower on those broad comprehension questions when compared to other question types. However, Reading First-eligible students score one percentage point higher on these items relative to non-Reading First students. These smaller, relative to vocabulary, but still positive gains are what we would expect if better decoding frees up attention for understanding the passage itself (Gough and Tunmer, 1986).

Column 9 of Table 3 interacts the treatment with the share of advanced phonics words in each question, using the CIERA decodability categories defined in Section 3.2, and again finds a positive and significant interaction for Reading First-eligible students. Reading the three interactions together, the gains are concentrated where decoding may bind the most. The vocabulary differential is two times the broad comprehension differential. Moreover, in column 9 of Table 3, the main Reading First effect becomes small once we account for the share of advanced phonics words. Moreover, once column 9 accounts for the share of advanced phonics words, the main Reading First coefficient measures the advantage on a question with none of these words. That advantage is about half the overall reading advantage in Panel A of column 5, and it is statistically insignificant in the narrow-bandwidth sample (column 10). The reading gains seem to be present almost entirely in decodable content. The concentration we find at the question level mirrors what the national evaluation of Reading First found at the skill level, where the program improved decoding among first graders (Gamse et al., 2008).

Overall, the reading results suggest that Reading First strengthened a foundational skill rather than performance on a particular exam. If so, we would expect the gains to appear on other tasks that require students to read. The math exam in column 3 of Table 3 allows us to examine this hypothesis. Consistent with our earlier math test score results, Reading First eligibility increased the share of math questions answered correctly by 2.3 percentage points, an effect comparable to that of reading. The composition of the two exams may explain why the math effect is comparable to the effect we find on reading. Many math questions are word problems, and a student who cannot parse the question will do worse on these assessments (Abedi and Lord, 2001; Fuchs et al., 2015). Reading skill is therefore an input on much of the math exam, so an improvement in reading should raise math scores as well. Panel B provides suggestive evidence that reading is a channel behind the math gains. The first row of Panel

B measures the Reading First advantage on a hypothetical question with no words. On math questions that require little reading, students in Reading First districts show no detectable advantage over their peers. The advantage emerges as questions get longer and the gains are concentrated on the questions that demand the most reading. This pattern is also consistent with recent evidence that targeted interventions in one subject can spill over into another (Xu et al., 2023). Note that the results presented in Panel B are difficult to reconcile with the alternative explanation that math scores rose simply because eligible districts received more resources independent of phonics or Science of Reading-based instruction. If resources alone were driving our findings, we would expect gains on all math questions, including those with little text to read.

6.2 Expenditures

Since we find positive effects of Reading First on student test scores, it is of interest to know what the funds were allocated towards. Figure 4 summarizes the allocation of Reading First-specific expenditures using our main difference-in-differences model from equation (4). Each point is the estimated effect of eligibility on Reading First spending in the indicated category per K–3 student during the program years, 2004 through 2011. The four panels form a hierarchy. The first reports total tracked expenditures, and each subsequent panel decomposes the last category of the panel before it, so each panel has its own dollar scale. Panel A shows that eligible districts spent \$131.47 more per K–3 student from their Reading First fund relative to ineligible districts. We find similar, if somewhat smaller, effects when we restrict the comparison to districts within five percentage points of the 2002 fail rate cutoff of 13.6 percent that determined eligibility. [10]

Panel B breaks up the total expenditures reported in Panel A, into its broad components. The largest share went to salary, wages, and benefits (\$72.50), followed by supplies and materials (\$46.78). Districts also spent modestly more on professional purchased services (\$8.35) and travel (\$2.68).[11] Given that supplies and materials are the second largest category, Panel C decomposes this spending into its individual components, namely, general supplies, testing materials, and books. The increase splits roughly evenly between books (\$23.90) and general supplies (\$22.43), with a small and statistically significant effect on testing materials (\$0.41). Panel D further disaggregates books into reading materials, library books, and textbooks. Reading materials account for nearly all of the book spending result, \$22.25 of the \$23.90 total. The effects on library books (\$0.31) and textbooks (\$1.34) are statistically significant but economically small. The limited response of library books and textbooks may not be surprising given the program's focus on teaching students how to read, which presumably includes existing books. Library books and textbooks would not become obsolete because a district joined

[10] Appendix Tables B9, B10, and B11 report the underlying estimates, along with pooled versions, the post-2011 period, and corresponding 2SLS estimates. After 2011, spending in every category falls to near zero, consistent with the end of federal funding.

[11] There is no significant effect on miscellaneous operating costs, which we omit from the figure.

Reading First. Therefore, these results are consistent with phonics-based materials responding through the reading materials category, which is likely where phonics-related curricula would be cataloged.

Another possibility is that the increase in reading materials expenditures was used to update existing reading materials independent of phonics instruction. Thus, the reading improvements we find may be due to updated or new materials independent of the curricular content. However, Appendix Table B9 shows that these expenditures were not sustained beyond 2011, when Reading First funding ended, yet the third-grade test score effects persist, consistent with lasting changes in instructional practice. Moreover, the durable nature of books indicates that schools may have also purchased phonics-related materials and used them well beyond the end of the Reading First program in 2011 or that funds used to normally purchase non-phonics-based materials were now allocated to resources backed by the Science of Reading.

We next ask whether how a district allocated its award predicts the size of its achievement gains. For each treated district, we compute the share of its total Reading First award spent on each broad expenditure category and group districts into the bottom 2.5 percent, middle 95 percent, and top 2.5 percent of the distribution of that share. Figure 5 plots difference-in-differences estimates of the effect on standardized reading scores for two categories — supplies and materials, and salary, wages, and benefits — separately by these bins. Districts in the top 2.5 percent of the materials share, which would contain phonics-based reading materials, exhibit treatment effects of roughly 0.30 standard deviations.

Districts that spent the largest share on salaries had the smallest effects (roughly 0.08 standard deviations), while districts that spent the smallest share on salaries saw much larger effects near 0.30. For the middle 95 percent, the estimates are roughly 0.16 to 0.17 standard deviations. The pattern is consistent with the hypothesis that instructional materials were an important input, but since it is conditional on spending by category, these estimates may also reflect unobserved differences across districts. For example, a district with strong existing teachers but outdated materials might opt to buy materials and might have improved more because of the pre-existing strong base of teachers rather than the book purchases. Thus, the effects of spending within a particular category is potentially endogenous. For this reason, we read this figure as corroboration of the expenditure results rather than as a causal ranking of spending types. Figure 5 suggests that Reading First produced benefits under a wide range of allocations, with somewhat larger gains in reading among districts at the top of the supplies and materials distribution.

The salary estimates also preview the comparison we explore in Section 6.4, where we show that discretionary revenue from local tax elections flows predominantly to payroll with no purchases of reading materials, and produces much smaller reading gains. That contrast provides further suggestive evidence that Reading First's payroll spending differed from a pure increase in teacher compensation. Even payroll-heavy awards arrived bundled with the program's curricular mandates, coaching, and professional development, which may have made these dollars effective beyond the wage and retention channels documented in the teacher-pay

literature (Loeb and Page, 2000; Hendricks, 2014; Biasi, 2021).

6.3 Teacher and School Characteristics

Another mechanism explaining the effect of Reading First, could be changes in the composition of campus teachers or students. For instance, if higher-quality teachers moved to Reading First districts post-2004, then the effects we find could be attributed to teacher-specific quality rather than the pedagogical components of Reading First. Ideally, one would construct teacher-specific value-added measures from linked student-teacher data before Reading First (Chetty et al., 2014b; Rothstein, 2010) and examine the mobility of teachers across districts after the program. Unfortunately, our Texas data does not allow us to link students to teachers through classroom identifiers before 2012 (after Reading First funds ended). As an alternative, we follow Foy (2024) and estimate campus value added using student-level data from 1995 through 2002, the pre-Reading First period. Because third grade is the first tested grade in Texas during this period, we do not observe a lagged own-subject reading score for third graders. We address this limitation by constructing a value-added measure that consists of regressing standardized third-grade reading scores on lagged attendance (e.g., second grade attendance), a student's third-grade math scores, and student demographics (indicators for race, gender, economic disadvantage, ESL status, gifted status, special education, and immigrant status).[12] We take the residuals from the first-stage regression and collapse them to campus-year means. These means serve as our raw campus value-added estimates. We apply a standard empirical Bayes shrinkage procedure to the value added estimates, but the results are not meaningfully different if the raw value added estimates are used instead (Kane and Staiger, 2008).

To examine whether Reading First shifted the composition of the teaching workforce, we assign each teacher a value-added score based on the campus where they were employed in 2002, the year used to construct the Reading First eligibility criteria. Every teacher at a given campus receives the same score, corresponding to the campus-level value added from one of the three models described above. In each year, we compute the average value added of the teachers currently at that campus by averaging the scores the set of teachers there had from their 2002 placements. Thus if there was no teacher turnover or movement, value added would stay the same. We construct these means for all classroom employees at a campus as well as separately for teachers who entered the campus, teachers who exited, and those who remained (incumbents). Figure A10 reports the results. Across all models, we find no evidence that higher value-added teachers moved into Reading First districts after the program began. If anything, we see evidence of lower pre-2003 value-added teachers present in Reading First districts. These findings suggest that the reading achievement effects we document are not driven by a reshuffling of higher-quality teachers across Reading First and non-Reading First

[12] Figure A10 also considers two alternatives. One, where we do not include the third-grade math score and another measure of teacher value added that regresses standardized fourth-grade reading scores on fourth-grade student demographics, third-grade attendance, and the student's third-grade reading score. In all cases, our results are similar.

districts after the program began. As another proxy for teacher quality or preparedness, we consider changes in the share of teachers with a college degree in Figure A11. We do not find any convincing evidence of changes in the composition of teacher degree attainment. One exception is that the 2SLS estimates suggest a decrease in the share of teachers with no degree and an increase in the share with a bachelor's degree, particularly among employees entering the classroom.

We next examine whether Reading First affected teacher staffing levels. Figure A12 reports estimates where the outcomes are various measures of teacher full-time equivalents (FTE) per student. We consider the overall teacher-to-student ratio, the rate of new and departing teachers, K-3 FTE per student, and K-3 Reading/ELA FTE per student. The difference-in-differences estimates are small and not statistically significant across all staffing measures, suggesting no consistent effect of Reading First eligibility on the number of teachers per student.

Lastly, Figure A13 examines whether Reading First affected teacher compensation by estimating the effect on salary per FTE (in thousands of dollars). Note that we cannot observe which schools or employees were allocated Reading First funds, since we only have Reading First funding data at the district level. Because Reading First funds were distributed across a subset of campuses within a district and likely reached only a subset of staff within those campuses, this analysis is naturally attenuated relative to what we might find if we could link Reading First spending to individual employees. We find suggestive evidence that Reading First increased salaries, although the difference-in-differences estimates are less precise. This increase is consistent with our earlier finding that districts spent Reading First dollars partly on salaries.

6.4 Differences from Other Exogenous Sources of Funding

A natural question is whether our Reading First estimates reflect the effects of additional resources per se or something specific about how those resources were used. Reading First funds came bundled with curricular mandates, coaching requirements, and professional development - all of which could independently drive improvements in student outcomes. To separate the pecuniary from the curricular channel, we compare the effects of Reading First to those of a more general revenue shock that has been increasingly leveraged in the school spending literature, the passage of a school district tax election (Baron, 2022; Abott et al., 2020; Cellini et al., 2010; Martorell et al., 2016). We follow past work and leverage Texas school districts' ability to petition voters to raise the operational tax rate. Unlike Reading First, these funds carry no programmatic requirements and can be spent at the district's discretion. If additional pecuniary investments alone drive the Reading First results, we would expect tax election windfalls to produce similar effects on reading outcomes. Given that we can examine what tax election dollars are also spent on, we compare these investments alongside the Reading First spending findings. Thus, if districts spend general revenue windfalls differently from how Reading First funds are allocated, then we can map these spending differences onto differences

in the improvement in test scores. Together, we can provide suggestive evidence on how leveraging different inputs can affect test scores.

We estimate a difference-in-differences model using data on Texas school district tax elections between 2000 and 2019 merged with Texas PEIMS financial data (discussed earlier).[13] Treated districts are those where a tax increase passed, with control districts those that had not yet passed one or never did. We use the year of the first successful tax increase as the treatment year. Figure C1 presents the event study for local per-student revenue following a successful tax election. Local revenue rises after passage, by roughly $321 per student on average over the post-election years. This figure also shows a rise of similar size in federal revenue per student in eligible districts as a result of Reading First.

In Figure C1, we also show the effects of tax elections on reading test scores relative to the effect of Reading First. Both point estimates in Panel C capture the average treatment effect across the 2000–2019 period and come from the same campus-year panel with campus and year fixed effects. We find that tax elections do not move test scores during our sample period. Passing an election changes reading scores by an insignificant 0.02 standard deviation in the same sample, whereas Reading First raises scores by roughly 0.12 standard deviations.

Table C1 examines how these discretionary dollars are spent. As expected, school revenue increases for the average district post-tax election, and most of that increase comes from local revenue sources. The remainder, reflects state aid, since many of the later elections were rate swaps. In sharp contrast to Reading First, discretionary revenue increases are predominantly allocated to salaries and wages, with no significant effect on supplies and materials. Table C1 also shows that when we decompose the expenditure effect further into the supplies subcategories, we find no increase in reading materials or general supplies - the categories that constituted a substantial portion of Reading First spending. Taken together, these findings highlight that unrestricted revenue increases and Reading First funds were allocated to fundamentally different school inputs. This evidence suggests that the literacy components attached to Reading First play an important role in achieving the gains we document.

7 Long-Run Effects

Thus far, we find that Reading First improved student test scores, and the mechanisms above suggest the program durably changed classroom instruction. This section asks whether the resulting achievement gains persisted as exposed students aged, first through grades four to eight and then into high school, college, and the labor market. As described in Section 3, the ERC data link students to high school completion records, postsecondary enrollment and degree completion from the Texas Higher Education Coordinating Board, the National Student Clearinghouse and quarterly earnings from the Texas Workforce Commission. We

[13] To account for the staggered timing of elections across school districts, we estimate the interaction-weighted estimator of Sun and Abraham (2021), which remains valid when treatment effects vary across election cohorts and over time.

estimate fixed-effects OLS models for test scores, high school, and college outcomes and fixed-effects Poisson models for employment and earnings. Figure 6 and Table B12 report the results from our difference-in-differences specification. The first column of Figure 6 (Panels A-C) examines K–12 outcomes, and the second column (Panels D-F) examines college and labor market outcomes. Each plotted coefficient comes from a separate specification, and for every outcome we report estimates from the full sample of districts and from districts within five percentage points of the eligibility threshold. Analogous 2SLS estimates are shown in Appendix Figure A14.

7.1 Persistence Through High School

We first examine whether the Reading First effect on reading test scores persists beyond third grade for an individual student. Panel A of Figure 6 reports our main difference-in-differences estimate for third-grade reading at the student level rather than the campus-year level, still controlling for campus and year fixed effects. The estimates are similar to the more aggregated results presented earlier, with eligibility increasing third-grade reading scores by roughly 0.1 standard deviations. Panel A also plots effects on reading scores through grade eight. The gains persist across grades, waning over time but remaining statistically and economically meaningful for all grades. Panel B repeats this exercise for math. The individual-level estimates are again similar to our earlier campus-level results. Panel C of Figure 6 reports the effects of Reading First on whether a student ever drops out, receives a standard diploma on time (within four years), and ever graduates from high school. Students eligible for Reading First were about two percentage points less likely to drop out, 2.9 percentage points more likely to receive a standard diploma on time, and 2.4 percentage points more likely to ever graduate from high school. The relevant estimates for standard on time diploma and high school graduation are slightly higher since they include all third grade cohorts, irrespective of whether they are observed into adulthood (as is the case for post-high school outcomes).

7.2 College and the Labor Market

We next examine the effects of Reading First on college, and labor market outcomes. Because these post-high school outcomes require observing students into adulthood, the college, employment, and wage estimates are restricted to third-grade cohorts through 2014. Panel D of Figure 6 shows that Reading First increased the probability of enrolling in college by about 2.1 percentage points. This is driven by graduation from four-year colleges, although the four-year estimate on its own is imprecise. The two-year enrollment estimates are negative and imprecise. We also find a 3.3 percentage point increase in the probability of graduating from college, which is consistent with the enrollment result and is likewise concentrated in four-year institutions. One interpretation is that improved literacy matters most at the four-year margin. Two-year colleges typically have less restrictive admissions than four-year institutions, so an intervention that improves academic skills may matter most for admissions at a four-year college. The neg-

ative two-year point estimates may also be an indication of substitution, whereby students who would otherwise have attended a two-year college may instead enroll in a four-year institution. However, these results are not as precise, and the bandwidth specifications are much smaller in magnitude.

Panels E and F of Figure 6 examine the likelihood of being in the labor force and associated annual wages at each age from 18 to 25. For these outcomes, we estimate fixed-effects Poisson models. One complication in the wage data is the interpretation of zeros which we address by imputing zero earnings for any individual who appears in the wage data at least once. The results in Figure 6 suggest that Reading First exposure raised earnings in adulthood. The likelihood of being employed is concentrated at age 18, fades by age 19, and is absent at later ages. Wages, by contrast, are higher at every age from 18 to 25, with the most precise and statistically significant estimates at ages 24 and 25, a timing that coincides with our college graduation effects. The difference-in-differences estimate in Appendix Table B12 implies that Reading First eligibility increased annual wages at age 25 by $554, or approximately 1.7 percent of the $32,500 mean among untreated students. The Poisson estimate plotted in the figure, 0.022 log points, is of the same size. [14] These magnitudes are consistent with other estimates of how early educational inputs translate into adult earnings. The effect of eligibility is comparable to assigning every student a one-standard-deviation better teacher for a single year (Chetty et al., 2014b).

In the 2SLS specification (Appendix Figure A14), the high school and college effects are close to the difference-in-differences estimates. However, the employment estimates in the 2SLS specification are slightly more negative and more precise, and the age-25 wage estimate is positive but smaller and imprecise. The funding groups explain the weaker wage estimate. Panels C and D of Appendix Figure A17 report the wage effects at age 25 separately for eligible districts in three groups of cumulative predicted funding (under $5 million, $5 to $15 million, and $15 million or more). Panel C evaluates each group's effect at its own mean funding, and Panel D rescales every estimate to the same $20 million. In levels, the wage effect rises from the low-spending group to the medium-spending group and then falls, so students in the highest-funded districts gained less than students in districts that received a fraction of the money. Per dollar, the wage effect declines from the low-spending group to the high-spending group, with the low-spending group's return roughly ten times that of the high-spending group. As with reading, most of the variation in predicted funding comes from the handful of very large districts, so the pooled per-dollar estimate is effectively the high-spending group's estimate. Since that group has the smallest per-dollar wage effect, the pooled estimate understates what the program did for wages in the low- and medium-spending districts. This is also where wages and reading differ. In Appendix Figure A17, the high-spending group's reading gains (Panel

[14] Two limitations of the age-25 earnings measure deserve emphasis. First, age 25 is early in the life cycle, and only captures a fraction of lifetime returns. Second, students who pursue further education (e.g., a master's degree) earn little while still enrolled. Given that we find that Reading First increases college enrollment, treated students may still be in school by the age of 25. This margin pushes the estimate toward zero, and the $554 figure likely understates the long-run impact of Reading First.

A) match the medium-spending group's, while its wage gains (Panel C) fall short of them. This pattern is consistent with wage effects driven more by participation in the program than by the amount of money districts received.

Given the wage gains, it is natural to ask why measured employment does not respond at later ages, and why the 2SLS estimate is, if anything, slightly negative. One possibility follows from the college results. Because Reading First increased four-year college graduation, some treated students may enroll in colleges out of state and remain outside Texas while employed, or graduate from a Texas university and take a job in another state. In either case, coding these students as having zero earnings would attenuate our estimates. Consistent with this, in Appendix Table B13 we find that eligible students are more likely than their peers to enroll in a college out of state, albeit the estimate is relatively smaller than the likelihood of going to college within Texas. Although not definitive, we take this as suggestive evidence for why wages rise while measured employment does not.

To understand the channels behind the earnings gain attributed to Reading First, Figure 7 presents a Gelbach decomposition (Gelbach, 2016), which simultaneously decomposes the total wage effect into the portions statistically accounted for by intermediate outcomes observed in the data, an accounting that, unlike adding controls one at a time, does not depend on the order in which the channels are entered. The decomposition should be read as a descriptive accounting of where the wage effect "lives" rather than as a set of independent causal mechanisms. An on-time high school diploma accounts for roughly \$153 of the total effect, enrollment in and completion of college (institution fixed effects) accounts for \$306, field of study (CIP fixed effects) accounts for \$110, and industry of employment (NAICS fixed effects) accounts for \$219. The component unexplained by these channels is negative (−\$234). This pattern suggests that a majority of the long-run earnings effect is attributable to improved educational attainment and the resulting occupational upgrading, rather than to unobserved channels. Appendix Table B14 reaches the same conclusion using expected-wage measures that assign each student the control-group mean wage for their college, field of study, or industry. This finding aligns with existing literature on the returns to higher education that identifies a wage premium from attending a public four year university (Kozakowski, 2023; Mountjoy, 2026; Smith et al., 2025). Both Kozakowski (2023) and Mountjoy (2026) show that the wage premium appears eight years after a student applies to college, approximately at ages 24-25, and grows over time. Our estimates of a wage premium are small in magnitude relative to the literature since we observe Reading First treated students only in the very early stages of their professional career (only through age 25).

Interestingly, the decomposition reveals that the smallest contributor is field of study. One hypothesis is that improved reading that is persistent through grade school may lead some to choose fields that may be more reading dependent. In Appendix Table C2, we examine this possibility directly. The Occupational Information Network (O*NET) is a Department of Labor database that scores occupations on the importance of specific skills, including reading comprehension. Using these reading-importance scores, we examine the extent to which Read-

ing First exposure altered the reading intensity of chosen majors, colleges, and broad industry of occupation (at the age of 25). We do not find any evidence that Reading First exposure shifted people into majors, colleges, or industries where reading is more important.

8 The Value of Reading First per Dollar Spent

Reading First was inexpensive relative to its effects. In fiscal terms, eligible districts received roughly $75 per K–3 student per year over the full post period, with an associated treatment effect of 0.11 standard deviations in reading score improvement. During the program (2004-2011), the $132 per student-year produced an effect of 0.07 standard deviations. Note that the full post period estimate benefits from persistence that required no additional spending. Under either estimate, the implied cost per 0.1 standard deviations is between $70 and $185 per student-year, compared to the roughly $3,000 per pupil per year that the school spending literature implies is typically required for a comparable gain (Jackson and Mackevicius, 2024).[15]

To formalize the comparison, we compute the Marginal Value of Public Funds (MVPF) of Reading First, following Hendren and Sprung-Keyser (2020) and using the difference-in-differences estimates rather than the pooled 2SLS slope. The MVPF is the ratio of beneficiaries' willingness to pay to the net cost borne by the government, where net cost is the program spending less the additional tax revenue generated by the resulting increase in earnings. We adopt the analysis of Jackson et al. (2016) for K–12 school spending effects, and use an annual earnings gain of $554 at age 25 against cumulative Reading First during-program spending (assuming three years of exposure) of roughly $397 per exposed student, in 2025 dollars. When assessed relative to baseline earnings of $32,500, the $554 Reading First induced wage premium represents a 1.7% increase or 0.04003% per Reading First dollar spend. We hold this earnings gain constant throughout the entire age-earnings profile, allow for a 0.5 percent annual real wage growth, and project the gain forward through age 65. We discount both benefits and costs back to the average age at which the money was spent, and Reading First students were exposed to the treatment (age 7). Under these assumptions, the additional tax revenue exceeds the initial outlay– the program pays for itself, and the MVPF is unbounded– starting at age 30. Appendix C.3 sets out the construction in full. Table C3 stops the earnings projection at successively later ages and reports the net cost to the government at each. Crediting Reading First with only the single year of earnings we observe, at age 25, and assuming the effect vanishes thereafter, the government recovers sixteen cents of each dollar spent, and the MVPF is 0.79. We note that this projection is conservative given that our method holds the proportional earnings effect constant, whereas Mountjoy (2026) finds that the earnings returns to attending a Texas public university more than double between ages 25 and 30.

[15] One caveat is that the federal grant likely understates Reading First's full cost of implementation. The program relied on existing teacher time, administrative capacity, and professional development budgets. Moreover, the state retained up to 20 percent of federal funds for program support. None of these costs appear in the district-level grant figures. Nonetheless, even if unmeasured implementation costs double or triple the per-student figure, Reading First remains relatively cost-effective compared to unconditional revenue.

9 Conclusion

Reading proficiency is an important predictor of later-in-life achievement. Although substantial resources have been allocated towards improving reading curricula in elementary schools, evidence of the long term benefits have been limited. We study a phonics based policy intervention that successfully improved reading scores among third graders in Texas, sustained those improvements through later grades, and had performance spillovers in math. Using administrative data from the Texas Education Agency coupled with predicted receipts of Reading First funds to eligible school districts, we can speak to both the educational benefits of the policy and its cost-effectiveness. Contrary to unconditional influxes of funding, Reading First included specific guidelines on how districts should allocate the money they received. We track those expenditures and show that they are spent to a large extent on reading materials and teacher salaries. At a direct fiscal cost of roughly \$70–\$185 per student-year, Reading First achieved an increase of 0.1 standard deviations in reading test scores, making it substantially more cost-effective than other interventions studied in the literature.

Our analysis of reading and math exam questions provides substantive evidence of the mechanism through which Reading First improved student performance. Reading First-eligible students performed better in longer, more complex questions than their non-Reading First-eligible peers. We are able to follow treated students through college and the early stages of their labor market participation. We find an increased probability of graduating from four year universities and a wage premium of \$554 at age 25. Overall, our work informs the debate over the effectiveness of early childhood intervention by showing that returns depend on how government funding is spent, and not only on how much is allocated. At a relatively low-cost, Reading First generated lasting effects that substantially improved both short term and long term outcomes for exposed students.

References

Abedi, Jamal and Carol Lord (2001) "The Language Factor in Mathematics Tests," *Applied Measurement in Education*, 14 (3), 219–234, 10.1207/S15324818AME1403_2.

Abott, Carolyn, Vladimir Kogan, Stéphane Lavertu, and Zachary Peskowitz (2020) "School district operational spending and student outcomes: Evidence from tax elections in seven states," *Journal of Public Economics*, 183, 104142.

Alesina, Alberto, Reza Baqir, and Caroline Hoxby (2004) "Political Jurisdictions in Heterogeneous Communities," *Journal of Political Economy*, 112 (2), 348–396, 10.1086/381474.

Algozzine, Bob, Chuang Wang, and Amy S Violette (2011) "Reexamining the relationship between academic achievement and social behavior," *Journal of positive behavior interventions*, 13 (1), 3–16.

Altonji, Joseph G., Lisa B. Kahn, and Jamin D. Speer (2016) "Cashier or consultant? Entry labor market conditions, field of study, and career success," *Journal of Labor Economics*, 34 (S1), S361–S401.

Álvarez Marinelli, Horacio, Samuel Berlinski, and Matías Busso (2024) "Remedial Education: Evidence from a Sequence of Experiments in Colombia," *Journal of Human Resources*, 59 (1), 141–174.

Ang, Desmond (2021) "The effects of police violence on inner-city students," *The Quarterly Journal of Economics*, 136 (1), 115–168.

Bacher-Hicks, Andrew, Joshua Goodman, and Christine Mulhern (2021) "Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time," *Journal of Public Economics*, 193, 104345.

Bacher-Hicks, Andrew and Cory Koedel (2023) "Estimation and Interpretation of Teacher Value Added in Research Applications," in *Handbook of the Economics of Education*, Vol. 6, 93–134.

Baicker, Katherine and Douglas Staiger (2005) "Fiscal Shenanigans, Targeted Federal Health Care Funds, and Patient Mortality," *Quarterly Journal of Economics*, 120 (1), 345–386, 10.1162/0033553053327461.

Bailey, Drew H., Greg J. Duncan, Flavio Cunha, Barbara R. Foorman, and David S. Yeager (2020) "Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions," *Psychological Science in the Public Interest*, 21 (2), 55–97, 10.1177/1529100620915848.

Baron, E Jason (2022) "School spending and student outcomes: Evidence from revenue limit elections in Wisconsin," *American Economic Journal: Economic Policy*, 14 (1), 1–39.

Biasi, Barbara (2021) "The Labor Market for Teachers under Different Pay Schemes," *American Economic Journal: Economic Policy*, 13 (3), 63–102.

Boser, Ulrich, Matthew Chingos, and Chelsea Straus (2015) "The hidden value of curriculum reform: Do states and districts receive the most bang for their curriculum buck?", technical report, Center for American Progress.

Cabral, Marika, Bokyung Kim, Maya Rossin-Slater, Molly Schnell, and Hannes Schwandt (2021) "Trauma at school: The impacts of shootings on students' human capital and economic outcomes," technical report, National Bureau of Economic Research.

Cascio, Elizabeth U., Nora Gordon, and Sarah Reber (2010) "Paying for Progress: Conditional Grants and the Desegregation of Southern Schools," *Quarterly Journal of Economics*, 125 (1), 445–482, 10.1162/qjec.2010.125.1.445.

Cascio, Elizabeth U. and Douglas O. Staiger (2012) "Knowledge, Tests, and Fadeout in Educational Interventions," 10.3386/w18038, NBER Working Paper No. 18038.

Castles, Anne, Kathleen Rastle, and Kate Nation (2018) "Ending the reading wars: Reading acquisition from novice to expert," *Psychological Science in the Public Interest*, 19 (1), 5–51.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein (2010) "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design," *Quarterly*

Journal of Economics, 125 (1), 215–261.

Chen, Jiafeng and Jonathan Roth (2024) "Logs with zeros? Some problems and solutions," *The Quarterly Journal of Economics*, 139 (2), 891–936.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011) "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *The Quarterly journal of economics*, 126 (4), 1593–1660.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a) "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104 (9), 2593–2632.

——— (2014b) "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104 (9), 2633–2679, 10.1257/aer.104.9.2633.

Clay, Marie M. (1991) *Becoming Literate: The Construction of Inner Control*, Portsmouth, NH: Heinemann.

Cox, Robynn, Jamein P. Cunningham, Alberto Ortega, and Kenneth Whaley (2026) "Black Lives: The High Cost of Segregation," *American Economic Journal: Economic Policy*, 18 (1), 1–34.

Cunha, Flavio and James J. Heckman (2007) "The Technology of Skill Formation," *American Economic Review*, 97 (2), 31–47, 10.1257/aer.97.2.31.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010) "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78 (3), 883–931, 10.3982/ECTA6551.

Currie, Janet and Jonathan Gruber (1996) "Health insurance eligibility, utilization of medical care, and child health," *The Quarterly Journal of Economics*, 111 (2), 431–466.

Currie, Janet, Eric A Hanushek, E Megan Kahn, Matthew Neidell, and Steven G Rivkin (2009) "Does pollution increase school absences?" *The Review of Economics and Statistics*, 91 (4), 682–694.

Dahl, Gordon B and Lance Lochner (2012) "The impact of family income on child achievement: Evidence from the earned income tax credit," *American Economic Review*, 102 (5), 1927–1956.

Davis, Cassandra R., Jevay Grooms, Alberto Ortega, Joaquin Alfredo-Angel Rubalcaba, and Edward Vargas (2021) "Distance Learning and Parental Mental Health During COVID-19," *Educational Researcher*, 50 (1), 61–64.

Dee, Thomas S. (2024) "Higher chronic absenteeism threatens academic recovery from the COVID-19 pandemic," *Proceedings of the National Academy of Sciences*, 121 (3), e2312249121.

Denning, Jeffrey T., Richard Murphy, and Felix Weinhardt (2023) "Class Rank and Long-Run Outcomes," *The Review of Economics and Statistics*, 105 (6), 1426–1441.

Di, Qian, Itai Kloog, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, and Joel Schwartz (2016) "Assessing PM $_{2.5}$ Exposures with High Spatiotemporal Resolution across the Continental United States," *Environmental Science & Technology*, 50 (9), 4712–4721, 10.1021/acs.est.5b06121.

Duque, Daniel (2026) "The Long-Run Impact of School Funding on Economic Outcomes," MUNI ECON Working Paper2026-02, Masaryk University.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth (2016) "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution," *American Economic Journal: Applied Economics*, 8 (4), 36–65, 10.1257/app.20150213.

Foy, Morgan (2024) "Teachers' Unions and Student Achievement: Evidence from Wisconsin," https://morganfoy.github.io/papers/Foy_JMP.pdf.

Fuchs, Lynn S., Douglas Fuchs, Donald L. Compton, Carol L. Hamlett, and Amber Y. Wang (2015) "Is Word-Problem Solving a Form of Text Comprehension?" *Scientific Studies of Reading*, 19 (3), 204–223, 10.1080/10888438.2015.1005745.

Gamse, Beth C., Robin T. Jacob, Megan Horst, Beth Boulay, and Fatih Unlu (2008) "Reading First Impact Study: Final Report," technical report, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Gelbach, Jonah B. (2016) "When Do Covariates Matter? And Which Ones, and How Much?" *Journal of Labor Economics*, 34 (2), 509–543, 10.1086/683668.

Goodman, Kenneth S. (1967) "Reading: A Psycholinguistic Guessing Game," *Journal of the Reading Specialist*, 6 (4), 126–135.

Gordon, Nora (2004) "Do Federal Grants Boost School Spending? Evidence from Title I," *Journal of Public Economics*, 88 (9–10), 1771–1792, 10.1016/S0047-2727(03)00137-3.

Gough, Philip B. and William E. Tunmer (1986) "Decoding, Reading, and Reading Disability," *Remedial and Special Education*, 7 (1), 6–10, 10.1177/074193258600700104.

Hendren, Nathaniel and Ben Sprung-Keyser (2020) "A Unified Welfare Analysis of Government Policies," *The Quarterly Journal of Economics*, 135 (3), 1209–1318, 10.1093/qje/qjaa006.

Hendricks, Matthew D. (2014) "Does It Pay to Pay Teachers More? Evidence from Texas," *Journal of Public Economics*, 109, 50–63.

Hezel Associates, LLC (2007) "An Evaluation of Texas Reading First Activities, Materials and Providers," technical report, Texas Education Agency, Syracuse, NY.

Hoffman, James V, Misty Sailors, and Elizabeth U Patterson (2002) "Decodable texts for beginning reading instruction: The year 2000 basals," *Journal of Literacy Research*, 34 (3), 269–298.

Hollingsworth, Alex, Jiafang Mike Huang, Ivan Rudik, and Nicholas J Sanders (2022) "A Thousand Cuts: Cumulative Lead Exposure Reduces Academic Achievement," *Journal of Human Resources*, 0222–12169R2.

Hoxby, Caroline M. (1996) "How Teachers' Unions Affect Education Production," *Quarterly Journal of Economics*, 111 (3), 671–718, 10.2307/2946669.

——— (2000) "Does Competition among Public Schools Benefit Students and Taxpayers?" *American Economic Review*, 90 (5), 1209–1238, 10.1257/aer.90.5.1209.

——— (2021) "Advanced Cognitive Skill Deserts in the United States: Their Likely Causes and Implications," *Brookings Papers on Economic Activity*, Spring, 317–351.

Jackson, C Kirabo, Rucker C Johnson, and Claudia Persico (2016) "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms," *The Quarterly Journal of Economics*, 131 (1), 157–218.

Jackson, C Kirabo and Claire L Mackevicius (2024) "What impacts can we expect from school spending policy? Evidence from evaluations in the United States," *American Economic Journal: Applied Economics*, 16 (1), 412–446.

Jacob, Brian A., Lars Lefgren, and David P. Sims (2010) "The Persistence of Teacher-Induced Learning," *Journal of Human Resources*, 45 (4), 915–943.

Jacqz, Irene (2022) "Toxic test scores: The impact of chemical releases on standardized test performance within U.S. schools," *Journal of Environmental Economics and Management*, 102628, 10.1016/j.jeem.2022.102628.

Kane, Thomas J and Douglas O Staiger (2008) "Estimating teacher impacts on student achievement: An experimental evaluation," technical report, National Bureau of Economic Research.

Klein, Alyson (2008) "House Panel Would Kill Reading First Funding," *Education Week*, June 19.

Knight, Brian (2002) "Endogenous Federal Grants and Crowd-out of State Government Spending: Theory and Evidence from the Federal Highway Aid Program," *American Economic Review*, 92 (1), 71–92, 10.1257/000282802760015612.

Kogan, Vladimir (2025) *No Adult Left Behind: How Politics Hijacks Education Policy and Hurts Kids*, Cambridge Studies in the Comparative Politics of Education, Cambridge: Cambridge University Press.

Kozakowski, Whitney (2023) "Are Four-Year Public Colleges Engines for Economic Mobility?

Evidence from Statewide Admissions Thresholds. EdWorkingPaper No. 23-727.," *Annenberg Institute for School Reform at Brown University*.

Krueger, Alan B (1999) "Experimental estimates of education production functions," *The quarterly journal of economics*, 114 (2), 497–532.

Lee, David S., Justin McCrary, Marcelo J. Moon, and Martin Weidner (2022) "Valid t-Ratio Inference for IV," *American Economic Review*, 112 (10), 3260–3290, 10.1257/aer.20211063.

Loeb, Susanna and Marianne E. Page (2000) "Examining the Link between Teacher Wages and Student Outcomes: The Importance of Alternative Labor Market Opportunities and Non-Pecuniary Variation," *Review of Economics and Statistics*, 82 (3), 393–408.

Machin, Stephen, Sandra McNally, and Martina Viarengo (2018) "Changing How Literacy Is Taught: Evidence on Synthetic Phonics," *American Economic Journal: Economic Policy*, 10 (2), 217–241, 10.1257/pol.20160514.

Marshall, John (2024) "Can Close Election Regression Discontinuity Designs Identify Effects of Winning Politician Characteristics?" *American Journal of Political Science*, 68 (2), 494–510, 10.1111/ajps.12741.

Martorell, Paco, Kevin Stange, and Isaac McFarlin (2016) "Investing in Schools: Capital Spending, Facility Conditions, and Student Achievement," *Journal of Public Economics*, 140, 13–29.

Mathews, Mitford M. (1966) *Teaching to Read, Historically Considered*, Chicago: University of Chicago Press.

McCallion, Gail (2005) "Reading First and Early Reading First: Background and Funding," Technical ReportRL31241, Congressional Research Service.

McIntosh, Kent, Robert H. Horner, David J. Chard, Crysta R. Dickey, and Drew H. Braun (2008) "Reading Skills and Function of Problem Behavior in Typical School Settings," *Journal of Special Education*, 42 (3), 131–147.

Morgan, Paul L, George Farkas, Paula A Tufis, and Rayne A Sperling (2008) "Are reading and behavior problems risk factors for each other?" *Journal of learning disabilities*, 41 (5), 417–436.

Mountjoy, Jack (2026) "Marginal Returns to Public Universities," *The Quarterly Journal of Economics*, 141 (1), 429–497.

National Reading Panel (2000) "Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction," Report of the National Reading Panel. National Institute of Child Health and Human Development.

NCES (2024) "The Nation's Report Card: 2024 Reading Assessment at Grades 4 and 8," national assessment of educational progress, U.S. Department of Education, Institute of Education Sciences, https://www.nationsreportcard.gov/reports/reading/2024/g4_8/, National Center for Education Statistics.

Novicoff, Sarah and Thomas S Dee (2026) "The achievement effects of scaling early literacy reforms," *Educational Evaluation and Policy Analysis*, 48 (2), 589–612.

Oates, Wallace E. (1999) "An Essay on Fiscal Federalism," *Journal of Economic Literature*, 37 (3), 1120–1149.

Ortega, Alberto and Tyler Ludwig (2022) "Immigrant English Proficiency, Children's Educational Performance, and Parental Involvement," *Review of Economics of the Household*, 1–27.

Paterson, Frances R. A. (2000) "The Politics of Phonics," *Journal of Curriculum and Supervision*, 15 (3), 179–211, ERIC Number: EJ603256.

Pearson, P. David (1989) "Reading the Whole-Language Movement," *The Elementary School Journal*, 90 (2), 231–241, 10.1086/461615.

Persico, Claudia L and Joanna Venator (2021) "The Effects of Local Industrial Pollution on Students and Schools," *Journal of Human Resources*, 56 (2), 406–445.

Rayner, Keith (1998) "Eye movements in reading and information processing: 20 years of

research," *Psychological Bulletin*, 124 (3), 372–422.

Reber, Sarah J (2010) "School desegregation and educational attainment for blacks," *Journal of Human resources*, 45 (4), 893–914.

Robelen, Erik W. (2011) "Budget Cuts Raise Questions About Federal Commitment to Literacy," *Education Week*, March 14.

Rothstein, Jesse (2010) "Teacher quality in educational production: Tracking, decay, and student achievement," *The Quarterly Journal of Economics*, 125 (1), 175–214.

Schwartz, Sarah (2022) "Which States Have Passed 'Science of Reading' Laws? What's in Them?" *Education Week*, https://www.edweek.org/teaching-learning/which-states-have-passed-science-of-reading-laws-whats-in-them/2022/07, Accessed: 2025-01-10.

Smith, Jonathan, Joshua Goodman, and Michael Hurwitz (2025) "The Economic Impact of Access to Public Four-Year Colleges," *Journal of Human Resources*, 10.3368/jhr.0324-13461R2, Published online before print, 4 February 2025.

Spencer, Noah (2024) "Comprehensive early literacy policy and the "Mississippi Miracle"," *Economics of Education Review*, 103, 102598.

Strunk, Katharine, Tanya Wright, Tara Kilbride, Qiong Zhu, Amy Cummings, Joanne West, Meg Turner, and Craig De Voto (2021) "Read by Grade Three: Year one report."

Sun, Liyang and Sarah Abraham (2021) "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 225 (2), 175–199, 10.1016/j.jeconom.2020.09.006.

Texas Education Agency (2024) "PEIMS Financial Data Downloads," https://tea.texas.gov/finance-and-grants/state-funding/state-funding-reports-and-data/peims-financial-data-downloads, Accessed August 1, 2024.

U.S. Department of Education (2002) "Guidance for the Reading First Program," technical report, U.S. Department of Education, Washington, DC.

——— (2011) "Reading First Implementation Study: 2008–09 Final Report," technical report, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, U.S. Department of Education, Washington, DC.

Wainer, Howard, Irina Grabovsky, and Daniel H. Robinson (2026) "On Education Miracles in General (and Those in Mississippi in Particular)," *Significance*, 23 (1), 32–36, 10.1093/jrssig/qmaf098.

Westall, John and Amy Cummings (2023) "The Effects of Early Literacy Policies on Student Achievement," *Available at SSRN 4427675*.

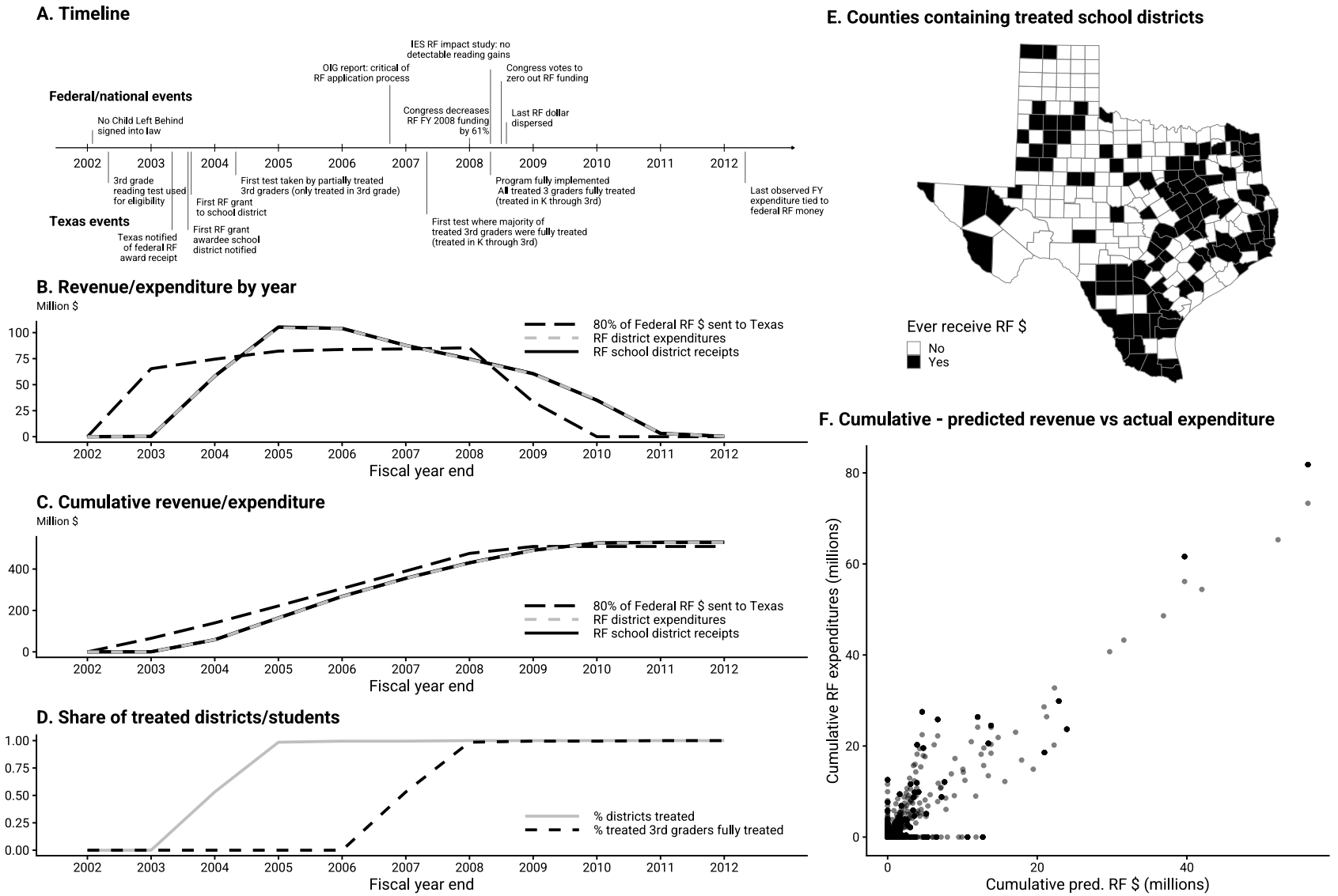
Whitehurst, Grover J (2009) "Don't forget curriculum," *Brown Center Letters on Education*, 12.

Xu, Zeyu, Umut Özek, Jesse Levin, and Dong Hoon Lee (2023) "Effects of Large-Scale Early Math Interventions on Student Outcomes: Evidence From Kentucky's Math Achievement Fund," *Educational Evaluation and Policy Analysis*, 01623737241303142.

Yoder, Maranna (2024) "Exhausted? The Effect of Diesel School Bus Emissions on Academic Achievement," technical report, Working Paper.

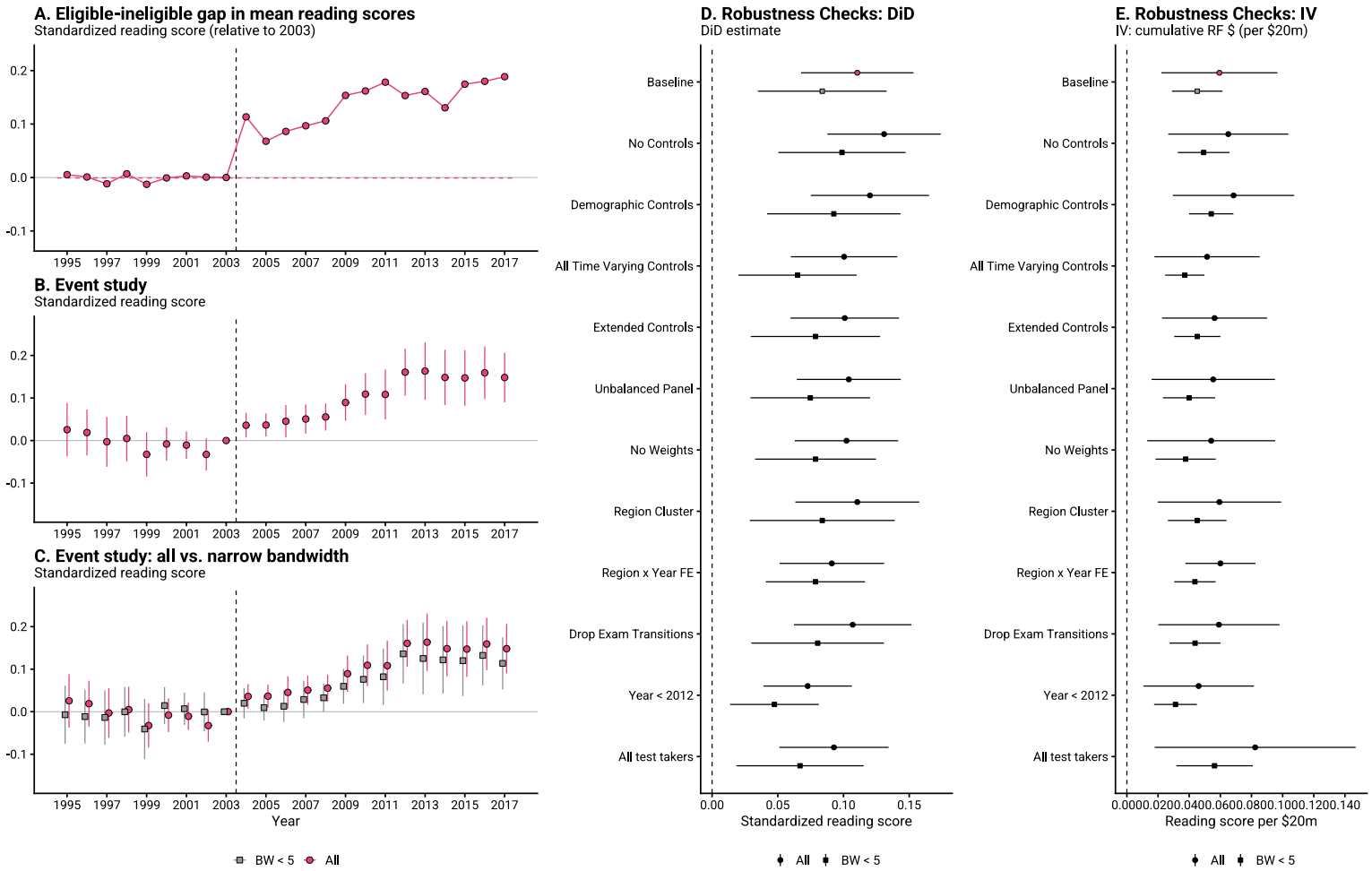
Zhong, Jiee (2026) "Early Grade Retention Harms Adult Earnings," *American Economic Journal: Applied Economics*.

Figure 1: Overview of spatial and temporal sources of variation



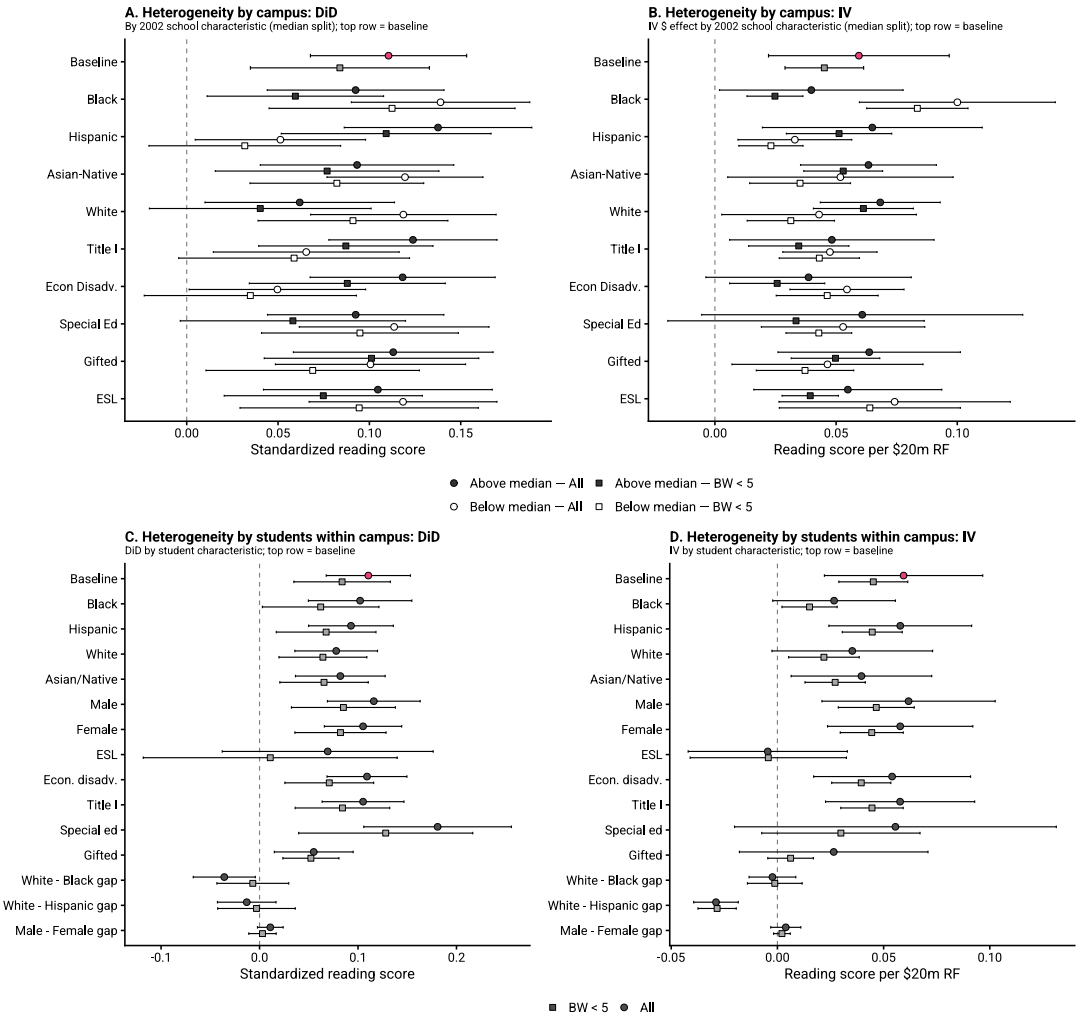
Notes: Panel A presents a timeline of Reading First (RF) in Texas alongside key federal events. Panel B (C) plots annual (cumulative) district RF receipts and expenditures by year, in millions of dollars. For every federal dollar received, Texas had to distribute at least 80% to school districts. Panel D displays the share of awarded districts that had received RF funds over time, along with the share of fully treated 3rd graders. Panel E maps the geographic distribution of RF-funded districts by Texas county. Panel F plots the relationship between predicted RF dollars received and actual RF expenditures.

Figure 2: Changes in third-grade reading by Reading First eligibility, relative to 2003



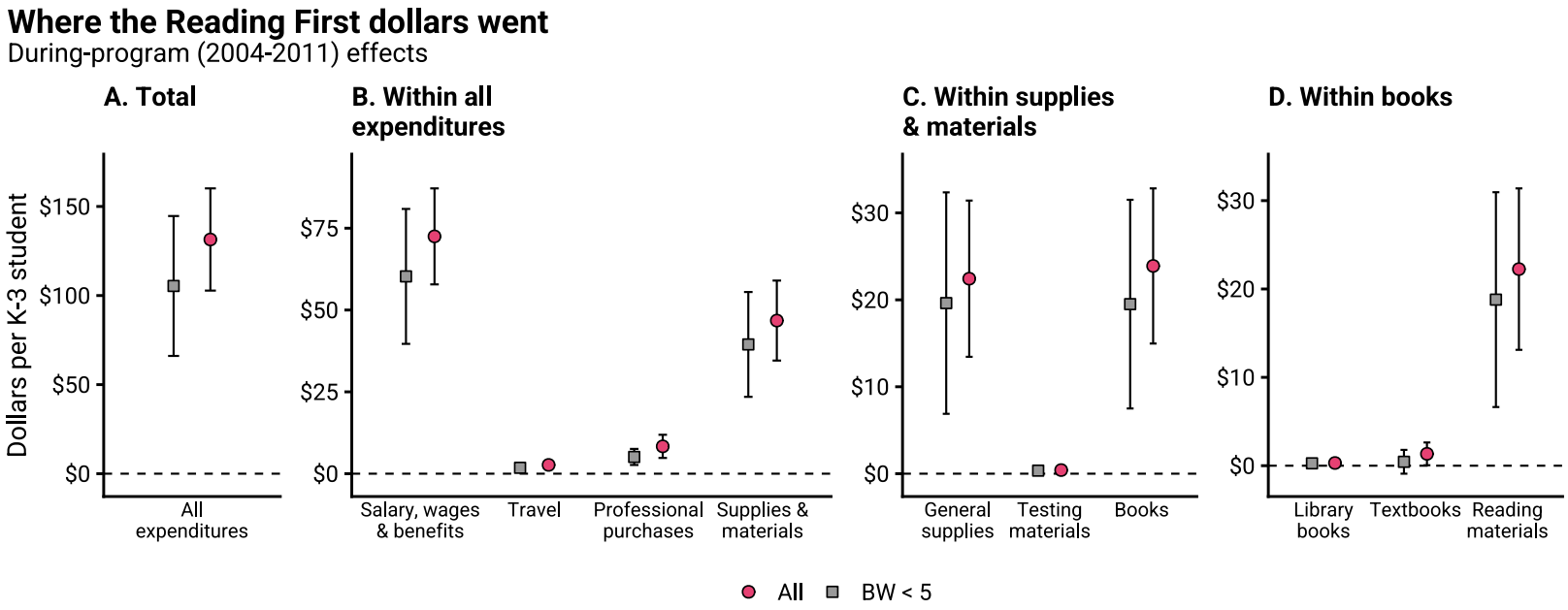
Notes: Panel A plots the raw difference in standardized third-grade reading scores between RF-eligible and non-eligible districts, relative to 2003. Panel B presents event-study estimates from equation (3) with 95% confidence intervals. Panel C adds a restricted sample of districts within five percentage points of the 13.6% fail rate eligibility threshold. Panel D reports specification checks for the difference-in-differences model and Panel E reports the corresponding checks for the 2SLS model (see Section 5.3 for details). All models include campus-by-grade and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects. Square markers restrict the sample to districts within five percentage points of the 13.6% fail rate threshold. Standard errors clustered at the district level. All samples are restricted to first-time test takers, except for the bottom row (all test takers) of Panels D and E. Panel E estimates are per $20 million of cumulative RF funding (2025 dollars).

Figure 3: Heterogeneity in the effect of Reading First eligibility on third-grade reading



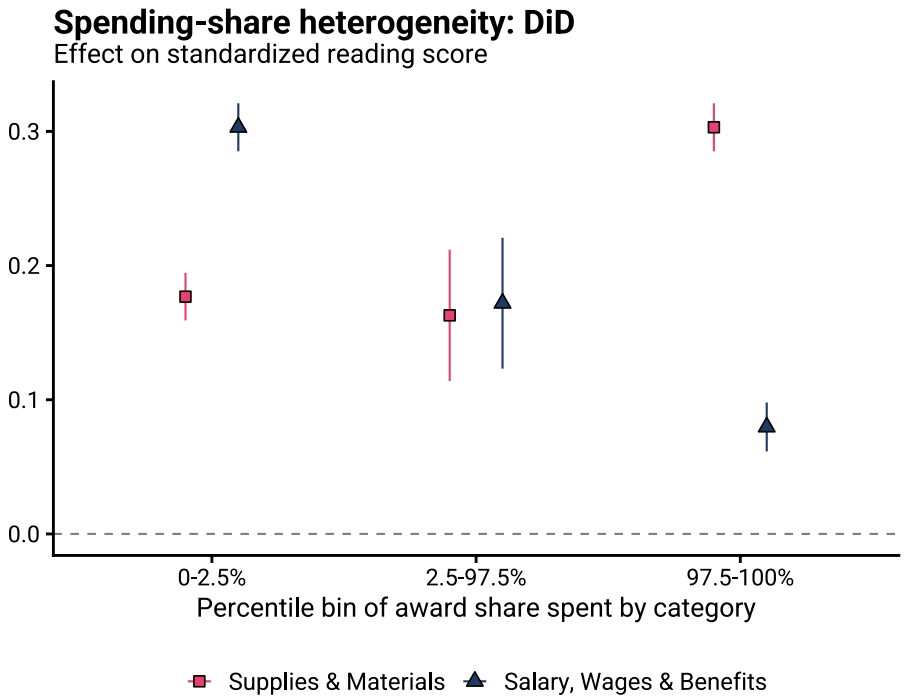
Notes: Panels A and B report split-sample estimates by 2002 campus characteristics (above/below median): Panel A reports difference-in-differences estimates and Panel B reports IV estimates per $20 million of cumulative RF funding (2025 dollars). Panels C and D report estimates by student characteristics within campuses, for the difference-in-differences and IV models, respectively. Square markers restrict the sample to districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure 4: Where the Reading First dollars went: during-program expenditures per K–3 student



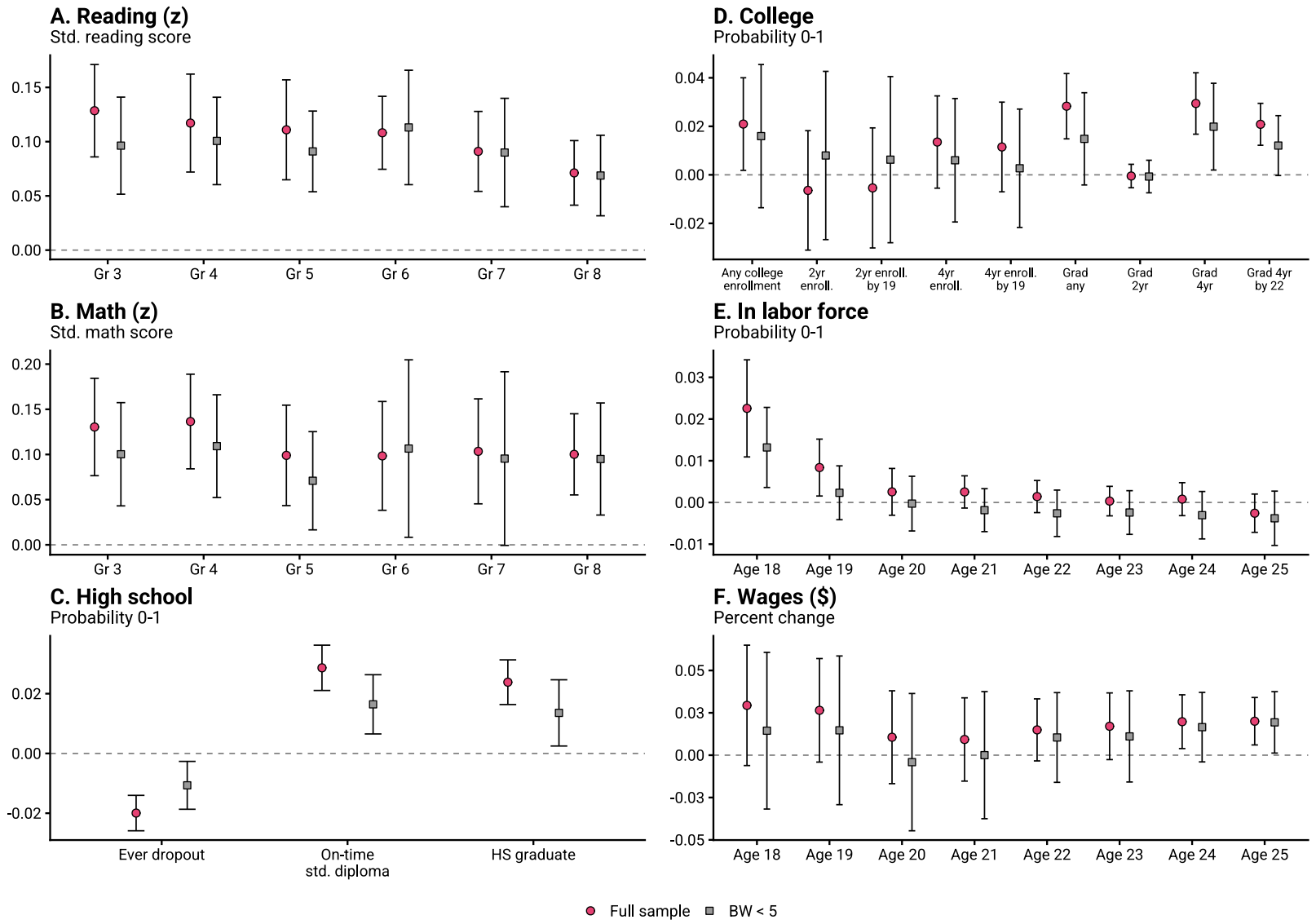
Notes: Each point reports the estimated effect of RF eligibility on the indicated RF expenditure category per K–3 student during the program period (2004–2011), in 2025 dollars. Circles use the full sample; squares restrict to districts within five percentage points of the 13.6% fail rate threshold. The four panels read left to right as a hierarchy and the last category of each panel is decomposed in the next: Panel A: total expenditures; Panel B: the breakdown of total expenditures into salary, wages & benefits; travel; professional purchases; supplies & materials; Panel C: the breakdown of supplies & materials into general supplies; testing materials; books; and Panel D: the breakdown of books into library books; textbooks; reading materials. Each panel has its own dollar scale. Whiskers represent 95% confidence intervals based on standard errors clustered at the district level.

Figure 5: Effect of Reading First eligibility on reading scores, by district spending composition



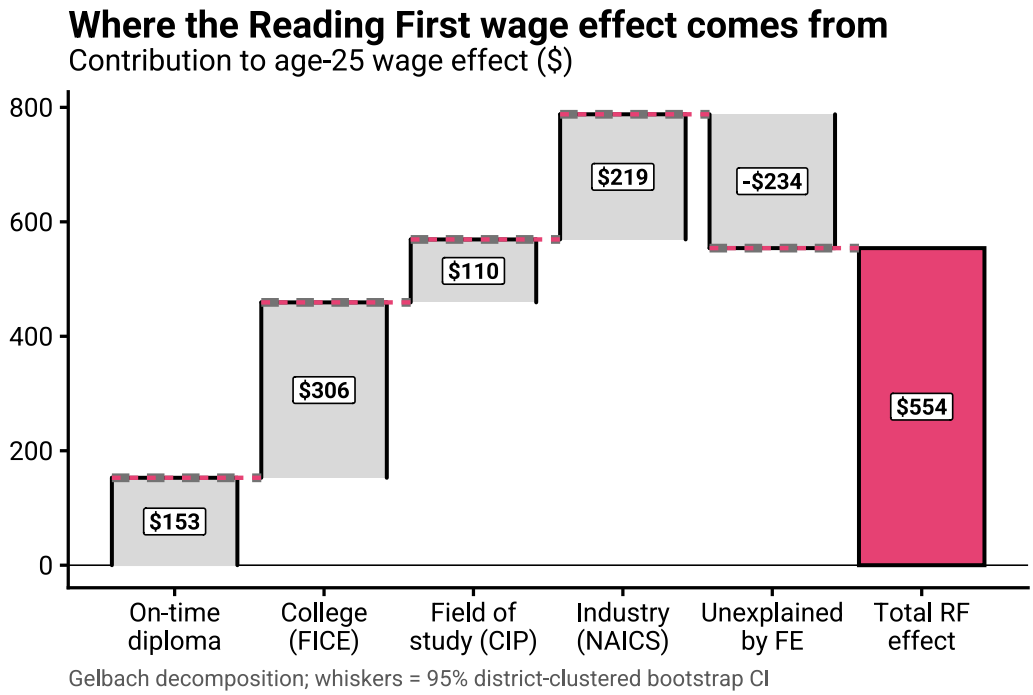
Notes: We split district level expenditures on Supplies & Materials and Salary, Wages & Benefits into three categories: 1) districts that spent 2.5 percent or less in those two categories, 2) districts that spent between 2.5 and 97.5 percent and 3) districts that spent over 97.5 percent. Each point displayed here is a difference-in-differences estimate of the effect that Reading First eligibility had on third-grade reading test scores. That effect is estimated separately for each of the three bins across the two expenditure categories. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure 6: Long-run effects of Reading First eligibility across the life course



Notes: Each panel reports estimates of the effect of RF eligibility on the indicated outcome: standardized reading and math scores, high school outcomes, college outcomes, labor force participation, and wages. Reading, math, high school, and college outcomes are estimated by OLS (in standard deviations or as linear probabilities); labor force participation and wages are estimated by a Poisson model (log points). Reading, math, and high school outcomes include all third-grade cohorts; college, labor force, and wage outcomes are restricted to cohorts through 2014. Within each panel, one series uses all districts and the other restricts the sample to districts within five percentage points of the 13.6% fail rate eligibility threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure 7: Where the Reading First wage effect comes from: Gelbach decomposition



Notes: The figure decomposes the total effect of RF eligibility on annual wages at age 25 ($554, rightmost bar) into the contribution of intermediate outcomes using a Gelbach decomposition: on-time diploma receipt ($153), college attended (FICE fixed effects, $306), field of study (CIP fixed effects, $110), industry of employment (NAICS fixed effects, $219), and the component unexplained by these fixed effects (−$234); all in 2025 dollars.

Table 1: Effect of Reading First eligibility on Reading First school district revenue

	Full sample	Bandwidth < 5
	(1)	(2)
A. Revenue per K–3 student, DiD		
1(RF Eligible) × 1(year>2003)	74.75***	58.85***
	(8.51)	(11.49)
B. Revenue per K–3 student, separate periods		
1(RF Eligible) × 1(2004–2011)	132.27***	105.51***
	(14.77)	(20.20)
1(RF Eligible) × 1(2012 and after)	1.57	0.01
	(1.47)	(1.69)
C. First stage, cumulative RF revenue		
Cumulative predicted RF $ ($ millions)	1.41***	1.40***
	(0.03)	(0.04)
First-stage F-stat	1,687.92	1,158.48
Observations	23,831	8,628
District FE	Yes	Yes
Year FE	Yes	Yes
Weights	K–3 students	K–3 students

Notes: The outcome in Panels A and B is Reading First-specific district revenue per K–3 student. Column (2) restricts to districts within predicted-funding bandwidth < 5. Panel A reports the pooled eligibility DiD and Panel B the split-period DiD (2004–2011 vs 2012 and after). Panel C is the first stage in cumulative levels: cumulative RF revenue regressed on cumulative predicted RF $ – the instrument used by the 2SLS panels – both in $ millions, with its first-stage F-statistic. All models include district and year fixed effects and are weighted by K–3 enrollment. District-clustered SEs in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Effect of Reading First eligibility on third-grade reading

	Full sample	Bandwidth < 5
	(1)	(2)
A. DiD		
1(RF Eligible) × 1(year>2003)	0.110***	0.084***
	(0.022)	(0.025)
B. DiD, separate periods		
1(RF Eligible) × 1(2004–2011)	0.072***	0.047***
	(0.017)	(0.017)
1(RF Eligible) × 1(2012 and after)	0.159***	0.130***
	(0.031)	(0.038)
C. 2SLS, cumulative RF $		
Cumulative RF $ per $20 million (2SLS)	0.059***	0.045***
	(0.019)	(0.008)
First-stage F-stat	1,711	863
Observations	60,619	34,240
Campus FE	Yes	Yes
Year FE	Yes	Yes

Notes: The outcome is standardized third-grade reading among first-time test takers. The sample is the balanced campus panel (campuses observed in every year 1995–2017), the paper's common school sample. Column (2) restricts to districts with predicted-funding bandwidth < 5. Panels report the pooled DiD, the split-period DiD, and 2SLS instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; the 2SLS coefficient is per $20 million. All models include campus and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects (the paper's baseline controls). District-clustered SEs in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

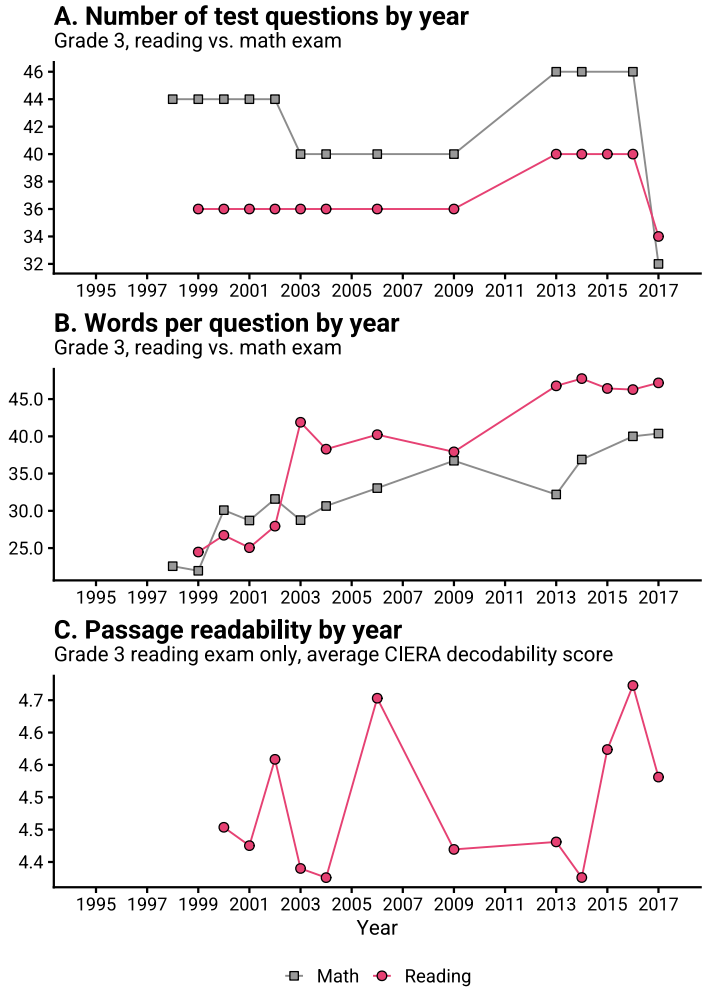
Table 3: Third-grade share of correct responses by Reading First eligibility

	Pooled		Math		Reading					
	All (1)	BW < 5 (2)	All (3)	BW < 5 (4)	All (5)	BW < 5 (6)	All (7)	BW < 5 (8)	All (9)	BW < 5 (10)
Panel A. Baseline DiD										
1(RF Eligible) × 1(year>2003)	2.308***	1.715***	2.259***	1.713**	2.323***	1.665***				
	(0.541)	(0.618)	(0.613)	(0.710)	(0.464)	(0.539)				
Observations	3,053,732	1,722,818	1,609,920	908,094	1,443,812	814,724				
Panel B. DiD × Word Count										
1(RF Eligible) × 1(year>2003)	1.104**	0.290	0.505	-0.593	1.299***	0.698				
	(0.537)	(0.593)	(0.605)	(0.660)	(0.443)	(0.499)				
Word Count	-0.153***	-0.156***	-0.203***	-0.213***	-0.135***	-0.134***				
	(0.003)	(0.004)	(0.006)	(0.007)	(0.002)	(0.003)				
1(RF Eligible) × 1(year>2003) × Word Count	0.030***	0.035***	0.048***	0.063***	0.023***	0.022***				
	(0.003)	(0.004)	(0.006)	(0.007)	(0.003)	(0.004)				
Observations	3,053,732	1,722,818	1,609,920	908,094	1,443,812	814,724				
Panel C. Reading Comprehension										
1(RF Eligible) × 1(year>2003)					2.070***	1.377**	1.882***	1.244**	1.159**	0.444
					(0.468)	(0.537)	(0.444)	(0.506)	(0.524)	(0.630)
Vocabulary					2.143***	1.939***				
					(0.104)	(0.136)				
1(RF Eligible) × 1(year>2003) × Vocabulary					2.174***	2.470***				
					(0.135)	(0.178)				
Broad Comprehension							-4.205***	-4.181***		
							(0.073)	(0.101)		
1(RF Eligible) × 1(year>2003) × Broad Comprehension							1.033***	0.984***		
							(0.122)	(0.178)		
Pct Advanced Phonics									0.062***	0.059***
									(0.003)	(0.004)
1(RF Eligible) × 1(year>2003) × Pct Advanced Phonics									0.077***	0.081***
									(0.006)	(0.008)
Observations					1,443,812	814,724	1,443,812	814,724	1,443,812	814,724
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The outcome is the share of questions answered correctly at the campus-year-grade-question level. The sample is the balanced campus panel (campuses observed in every year 1995–2017), the paper's common school sample. Columns (1)–(2) pool math and reading exams; columns (3)–(4) report results for math only; columns (5)–(10) report results for reading exams. Within each outcome, the first column uses the full sample and the second restricts to districts with predicted-funding bandwidth <5. Panel A: baseline difference-in-differences. Panel B: interacts the treatment with the word count of each question. Panel C (reading exams only): columns (5)–(6) interact with an indicator for vocabulary questions, columns (7)–(8) with an indicator for broad comprehension questions, and columns (9)–(10) with the share of advanced phonics words. All specifications include campus and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects (the paper's baseline controls). District-clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

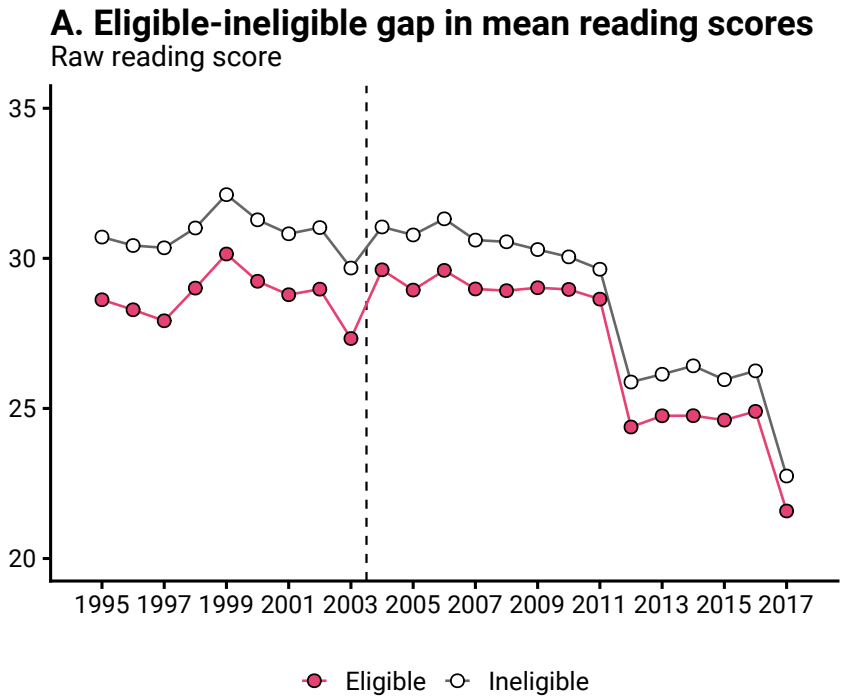
Appendix

Figure A1: Test-question trends for third-grade exams by year



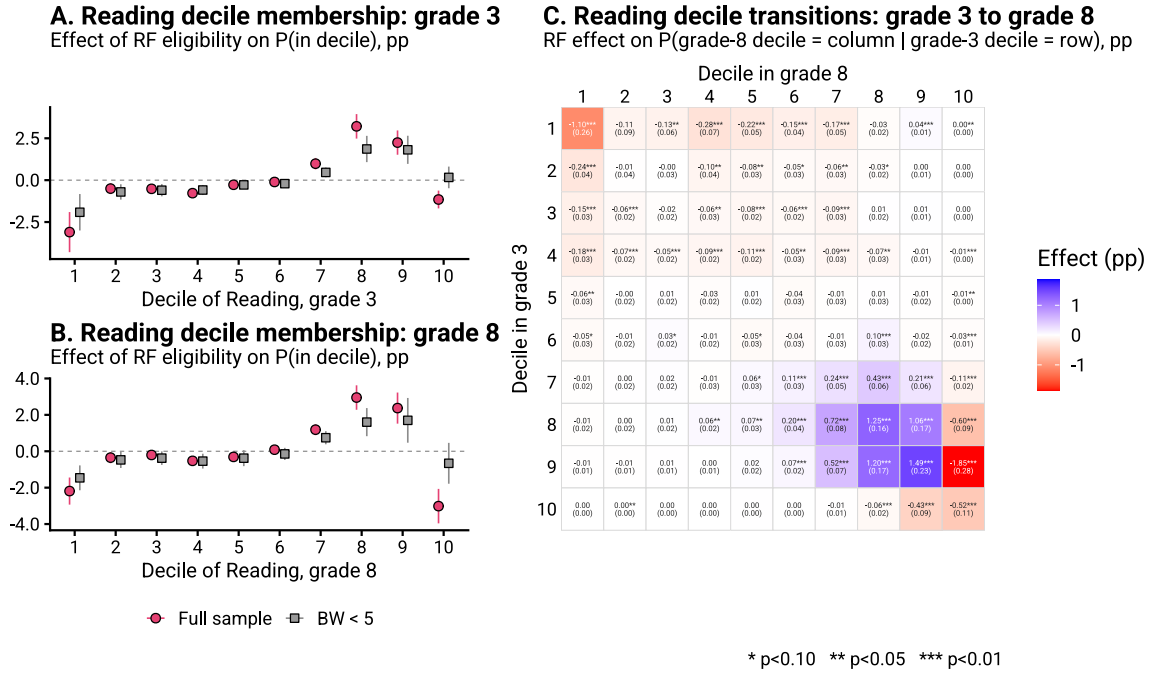
Notes: Panel A plots the number of released third-grade reading and math exam questions available by year. Panel B plots the average number of words per question for reading and math exams. Panel C shows the average CIERA-style decodability score for reading exams. Exams are drawn from the TAAS (1995–2002), TAKS (2003–2011), and STAAR (2012–present) testing programs. Released reading exams for 1995-1998, 2005, 2007, 2011, and 2012 could not be obtained from any source, and the 2008, and 2010 releases were practice exams rather than administered exams.

Figure A2: Trend in third-grade reading by Reading First eligibility



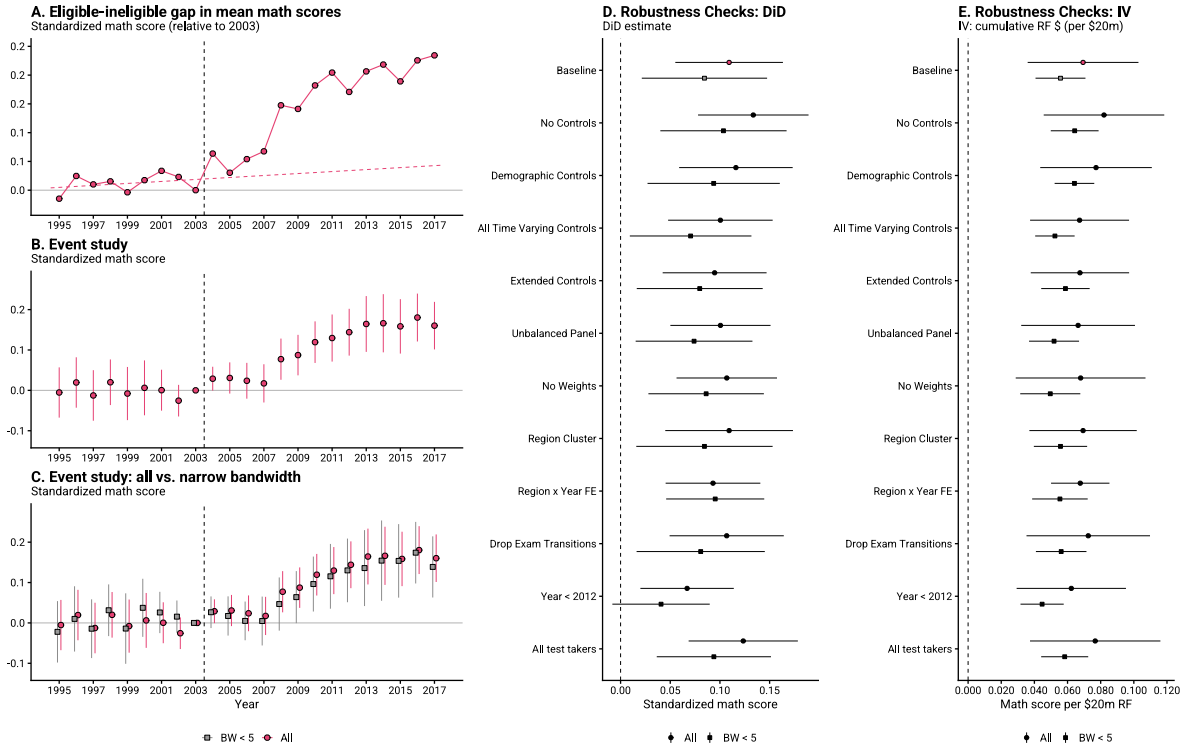
Notes: Mean raw (unstandardized) third-grade reading scores over time for Reading First-eligible and non-eligible districts: the unadjusted levels behind Panel A of Figure 2.

Figure A3: Effect of Reading First eligibility on reading decile membership and transitions, grades 3–8



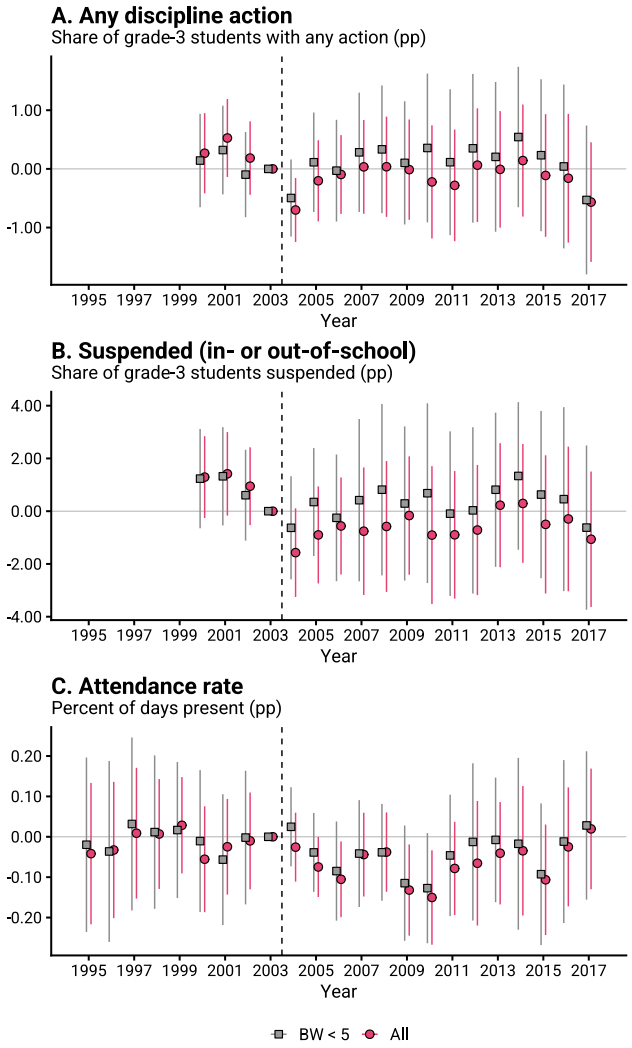
Notes: Panels A and B report difference-in-differences estimates of the effect of RF eligibility on grade-3 and grade-8 reading decile membership (percentage points); circles use the full sample and squares restrict to districts within five percentage points of the 13.6% fail rate threshold. Panel C reports the effect on each grade-3-to-grade-8 decile transition (row = grade-3 decile, column = grade-8 decile). Standard errors are clustered at the district level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A4: Effect of Reading First eligibility on third-grade math scores



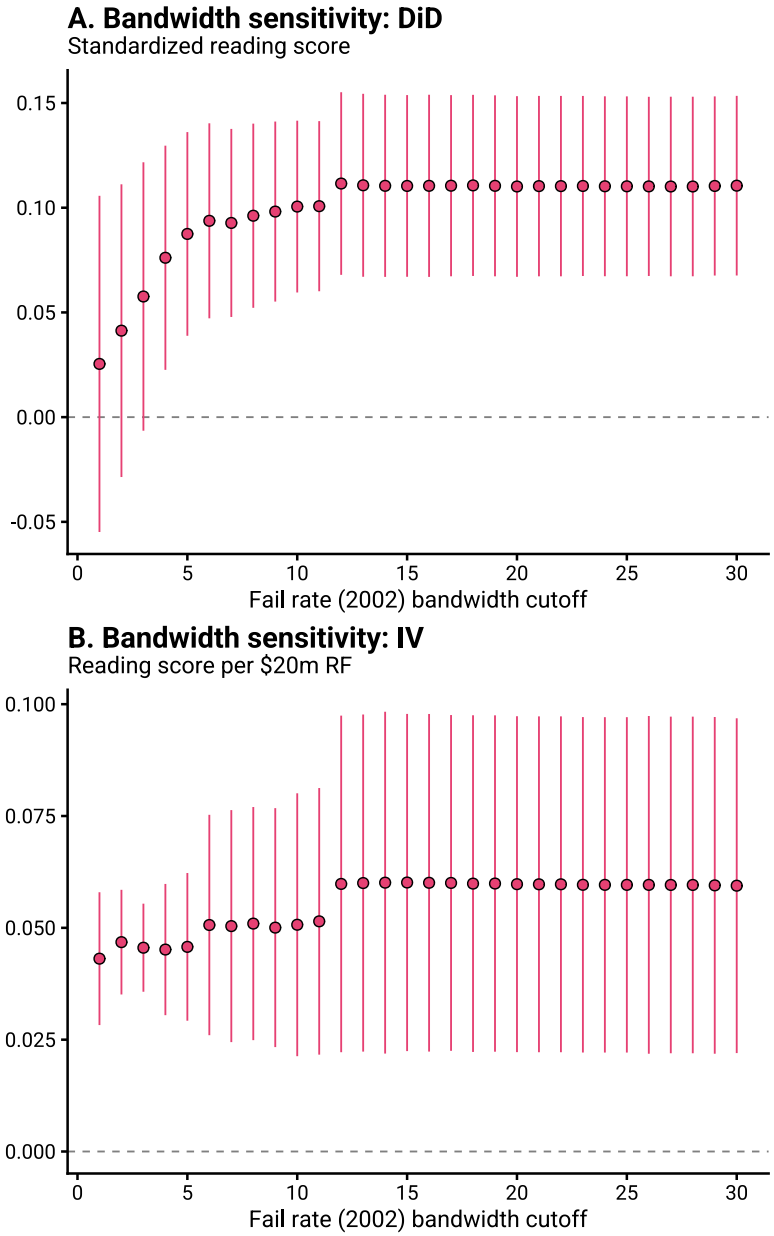
Notes: Math analogue of Figure 2. Panel A plots the raw difference in standardized third-grade math scores between RF-eligible and non-eligible districts, relative to 2003. Panel B presents event-study estimates with 95% confidence intervals, and Panel C compares the full sample with districts within five percentage points of the 13.6% fail rate eligibility threshold. Panels D and E report specification checks for the difference-in-differences and 2SLS models respectively. All models include campus-by-grade and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects. Square markers restrict the sample to districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level. All samples are restricted to first-time test takers, except for the bottom row (all test takers) of Panels D and E.

Figure A5: Event studies for behavioral outcomes



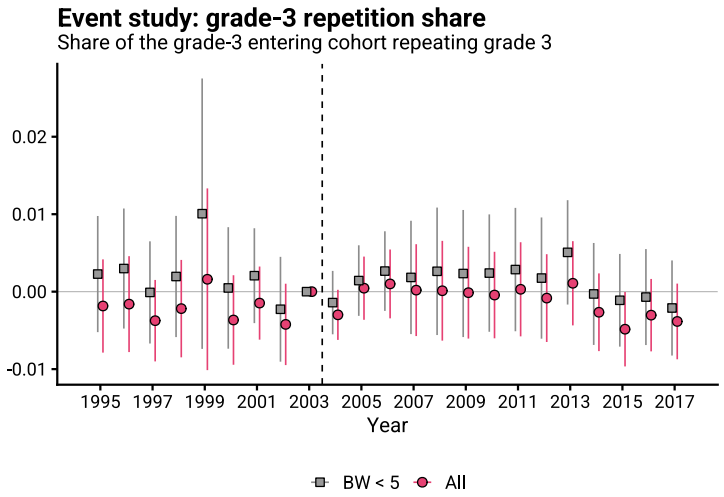
Notes: Event-study estimates (relative to 2003) of the effect of RF eligibility on three behavioral outcomes: the share of students with any disciplinary action (Panel A), the share of in or out of school suspensions(Panel B), and the attendance rate (Panel C). Attendance data are available for the entire sample period, while disciplinary records begin in 2000. Circles use the full sample; squares restrict to districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A6: Bandwidth sensitivity of the DiD and IV estimates



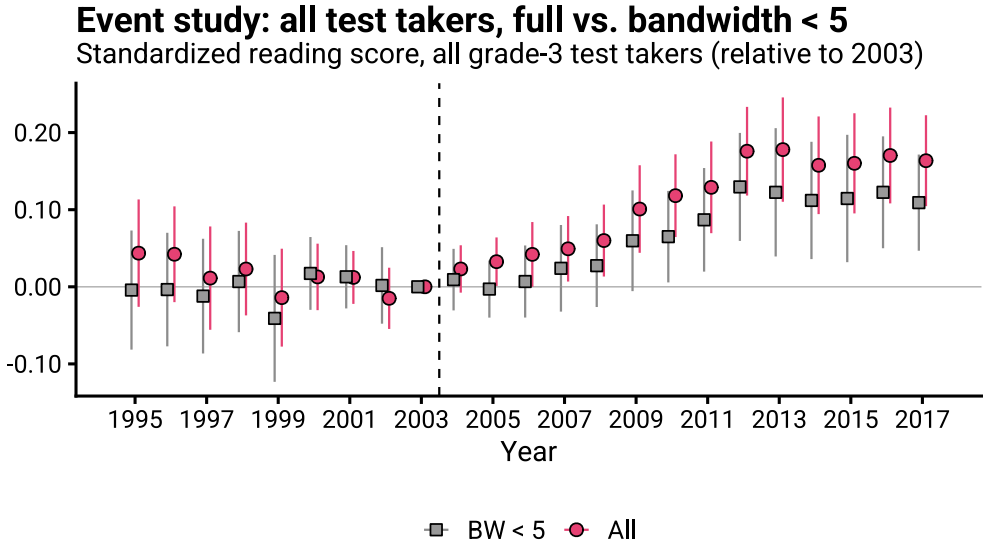
Notes: Each point reports the estimated effect on standardized third-grade reading scores, restricting the sample to districts within the indicated bandwidth (in percentage points) of the 13.6% fail rate eligibility threshold. Panel A reports the pooled difference-in-differences estimate of the effect of RF eligibility; Panel B reports the 2SLS estimate of the effect per $20 million of cumulative RF funding (2025 dollars). All models include campus-by-grade and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A7: Event study: grade-3 repetition share



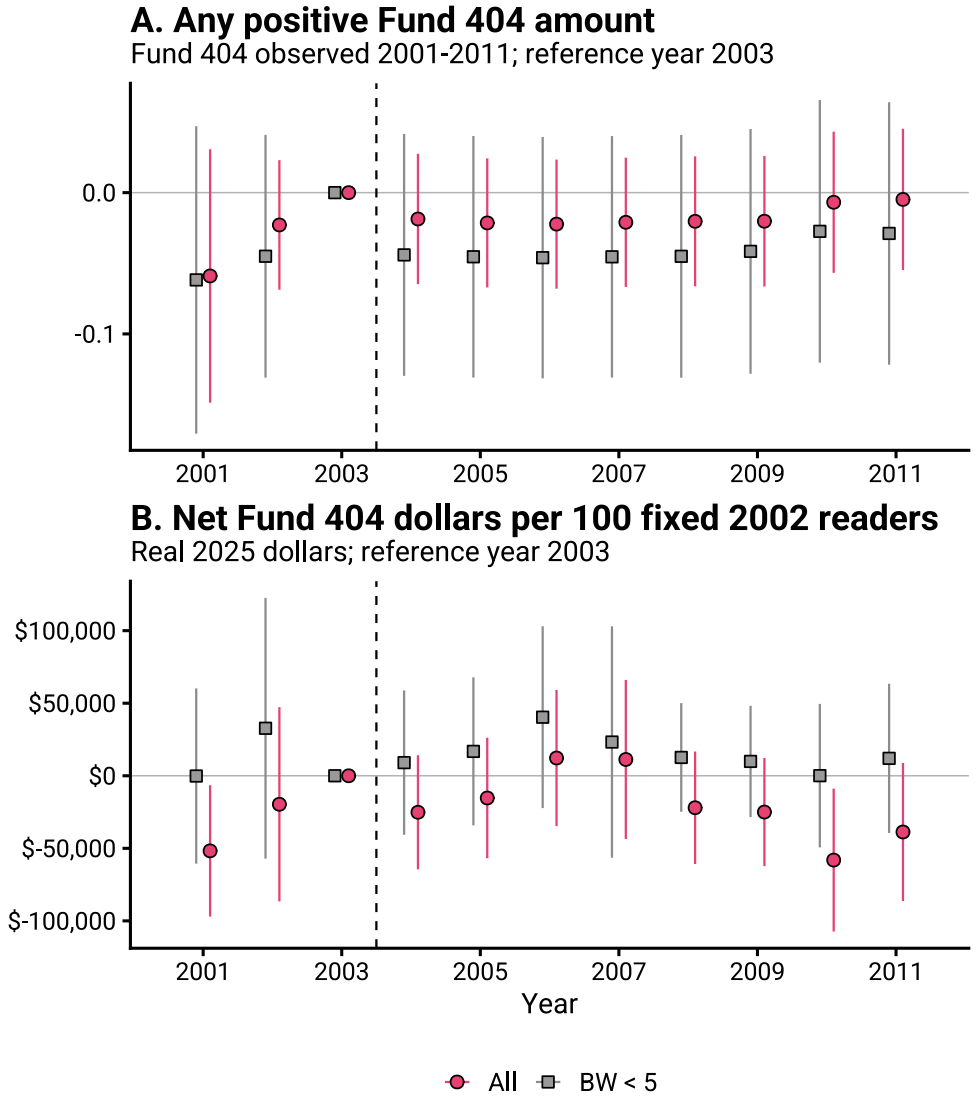
Notes: Event-study estimates of the effect of RF eligibility on the share of the third-grade entering cohort repeating grade 3, 1995–2017. Circles use the full sample; squares restrict to districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A8: Event study among all test takers: full sample vs. bandwidth < 5



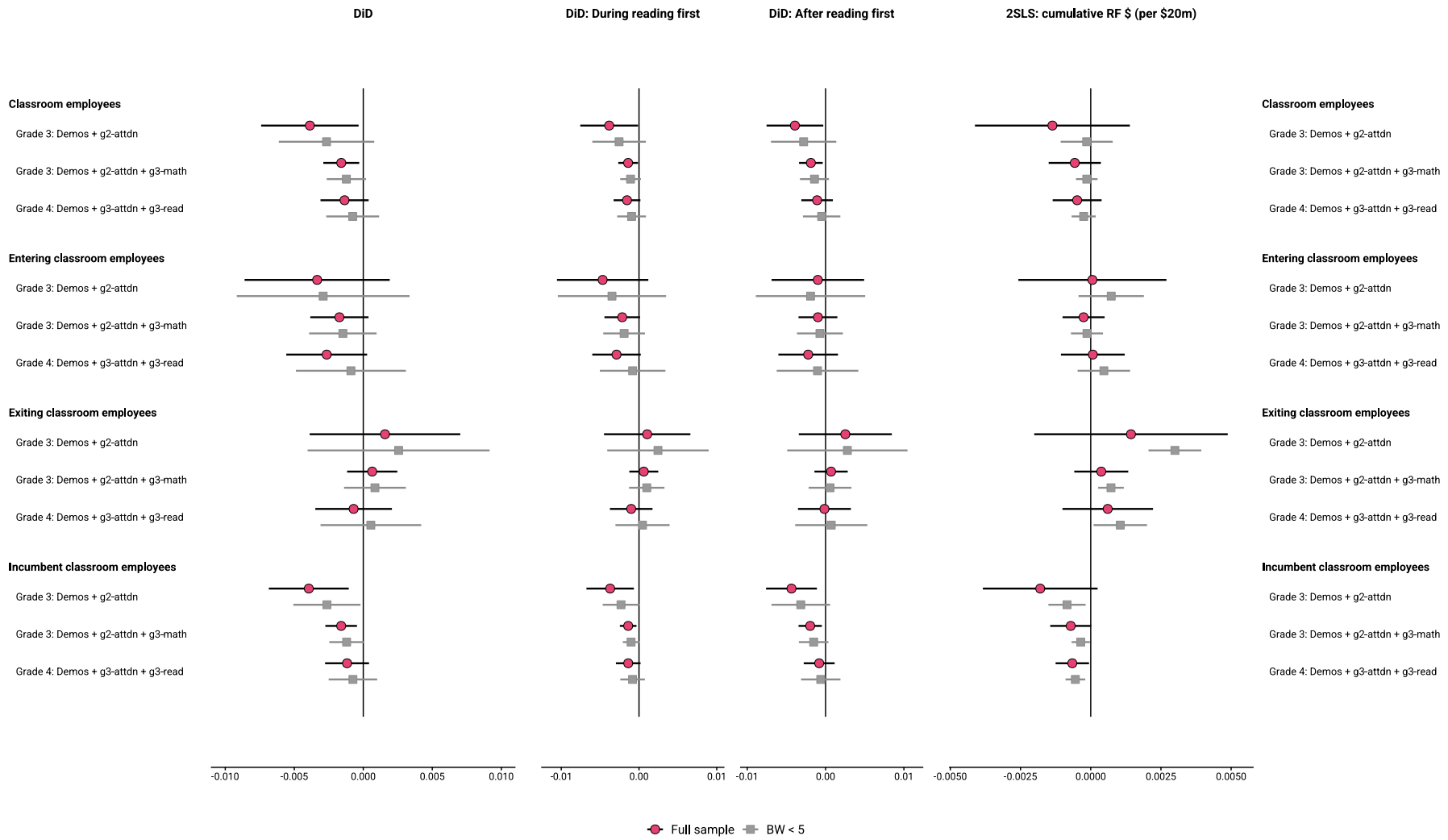
Notes: Event-study estimates of the effect of RF eligibility on standardized third-grade reading scores among all grade-3 test takers (relative to 2003), for the full sample and for districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level. This figure accompanies Table B3.

Figure A9: Student Success Initiative funding event studies



Notes: Event-study estimates of the effect of RF eligibility on the Student Success Initiative margins of Table B7: any positive Fund 404 amount and net Fund 404 dollars per 100 fixed 2002 grade-3 reading test takers (real 2025 dollars). Fund 404 is TEA's dedicated SSI fund. Estimates are relative to 2003, dating the mechanisms to SSI's 2003 onset rather than RF's 2004 outcome timing. Circles use the full sample; squares restrict to districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level. This figure accompanies Table B7.

Figure A10: Changes in average teacher value added by campus



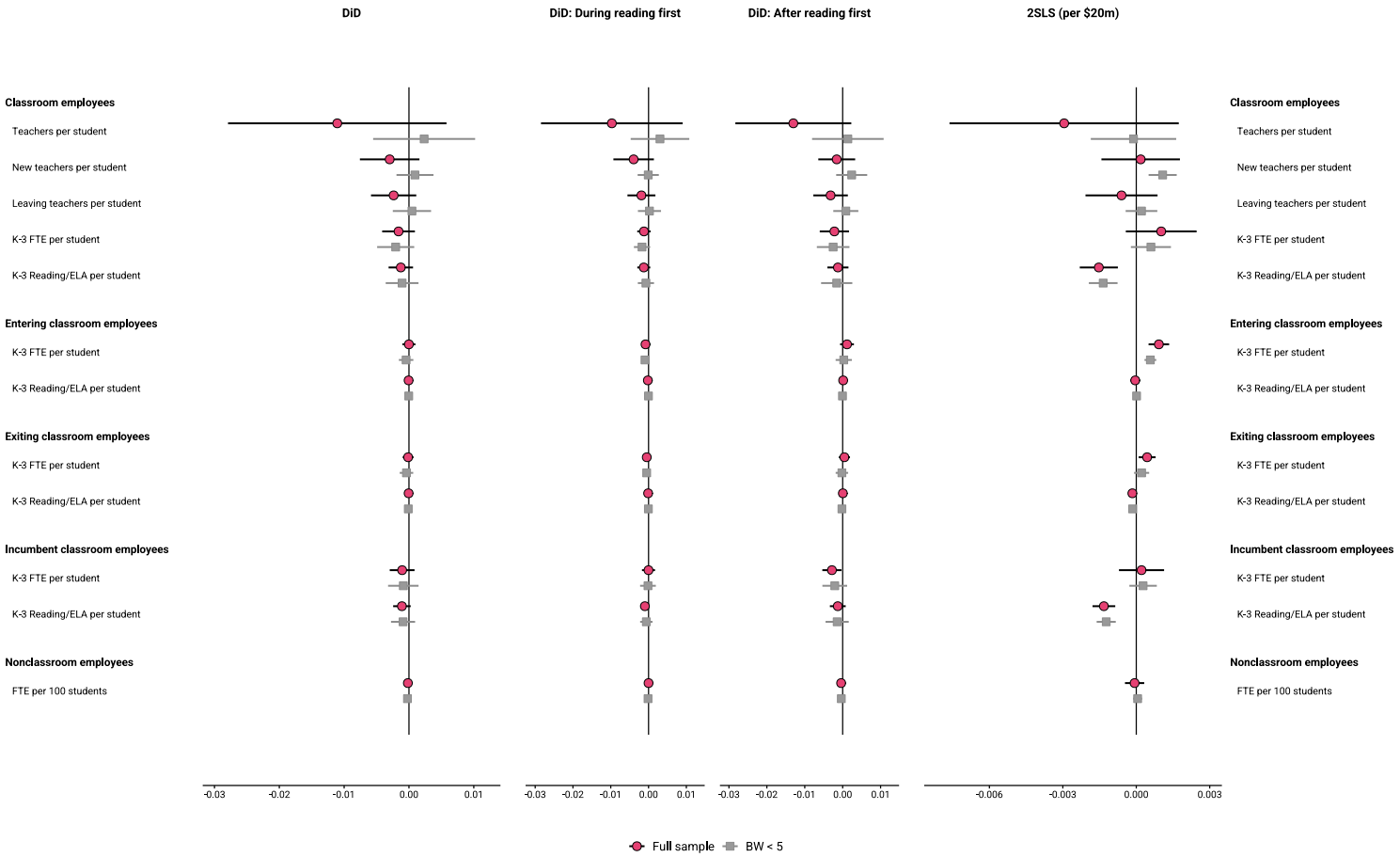
Notes: Each panel reports difference-in-differences, separate-period, and 2SLS estimates of the effect of RF eligibility on campus-level average pre-period employee value added. Value added is estimated using student-level data from 1995–2002 under three specifications (see Section 6.3). Teachers are assigned the value-added score of their 2002 campus and followed over time. Results are shown for all classroom employees, entering teachers, exiting teachers, and incumbents. Within each row, the pink circle uses the full sample and the grey square directly below restricts the sample to districts within five percentage points of the 13.6% fail rate eligibility threshold (BW < 5). Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A11: Changes in the share of teachers with degrees by campus



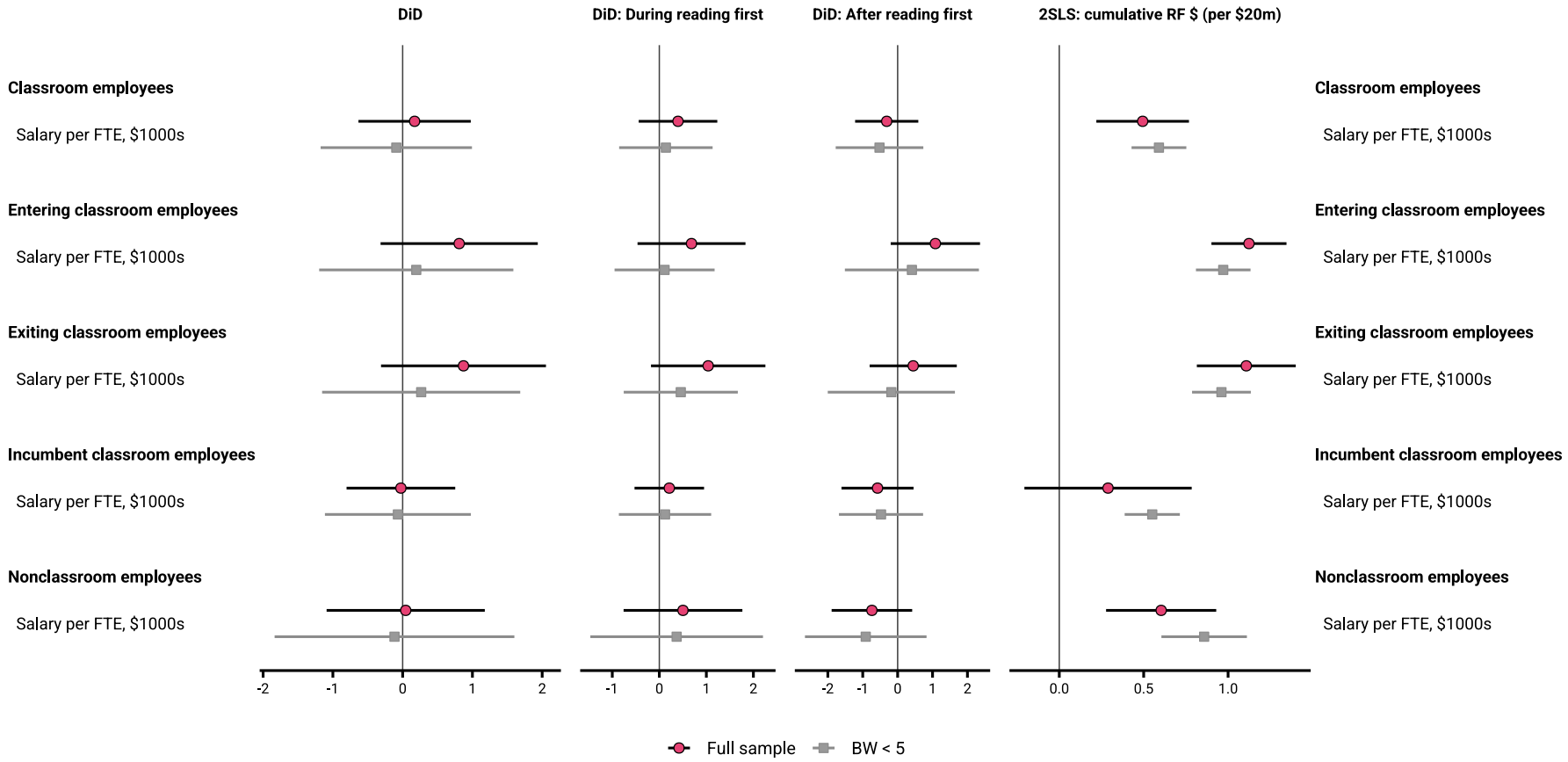
Notes: Each panel reports difference-in-differences, separate-period, and 2SLS estimates of the effect of RF eligibility on the share of classroom employees by degree attainment, for all classroom employees, entering, exiting, and incumbent employees. Within each row, the pink circle uses the full sample and the grey square directly below restricts the sample to districts within five percentage points of the 13.6% fail rate eligibility threshold (BW < 5). Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A12: Changes in employees by campus



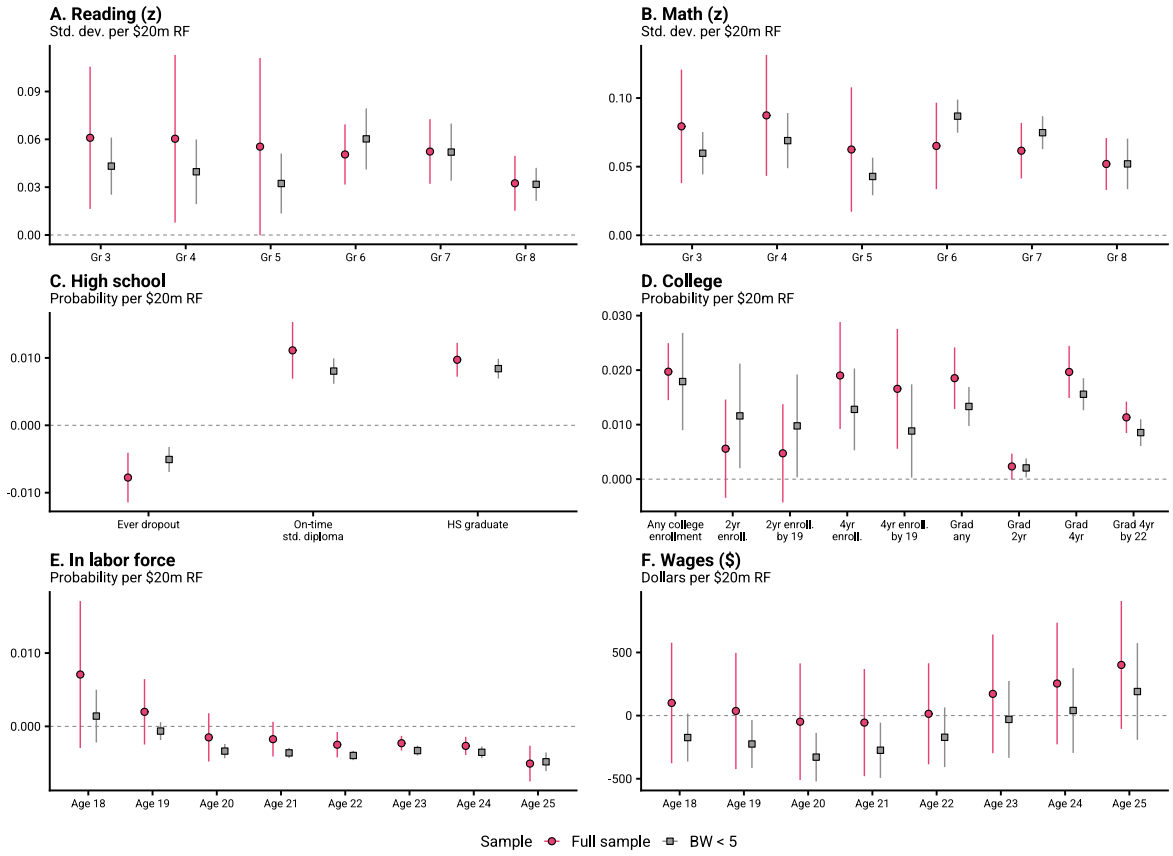
Notes: Each panel reports difference-in-differences, separate-period, and 2SLS estimates of the effect of RF eligibility on various measures of teacher full-time equivalents (FTE) per student. Outcomes include the overall teacher-to-student ratio, entering and exiting teachers, K–3 FTE per student, K–3 Reading/ELA FTE per student, and nonclassroom employee FTE per 100 students. Within each row, the pink circle uses the full sample and the grey square directly below restricts the sample to districts within five percentage points of the 13.6% fail rate eligibility threshold (BW < 5). Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A13: Changes in campus salary per FTE



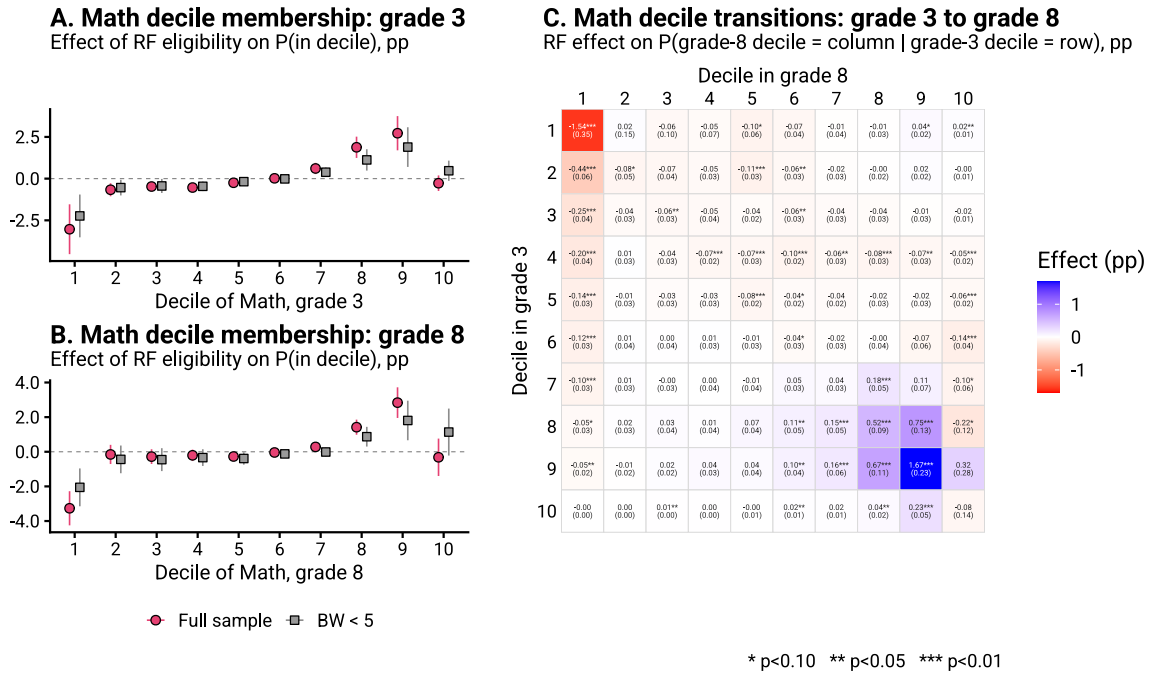
Notes: Each panel reports difference-in-differences, separate-period, and 2SLS estimates of the effect of RF eligibility on salary per FTE (in thousands of dollars). Results are shown separately for all classroom employees, entering, exiting, incumbents, and nonclassroom employees. Within each row, the pink circle uses the full sample and the grey square directly below restricts the sample to districts within five percentage points of the 13.6% fail rate eligibility threshold (BW < 5). Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A14: 2SLS estimates across the life course, per $20 million of cumulative Reading First funding (2025 dollars)



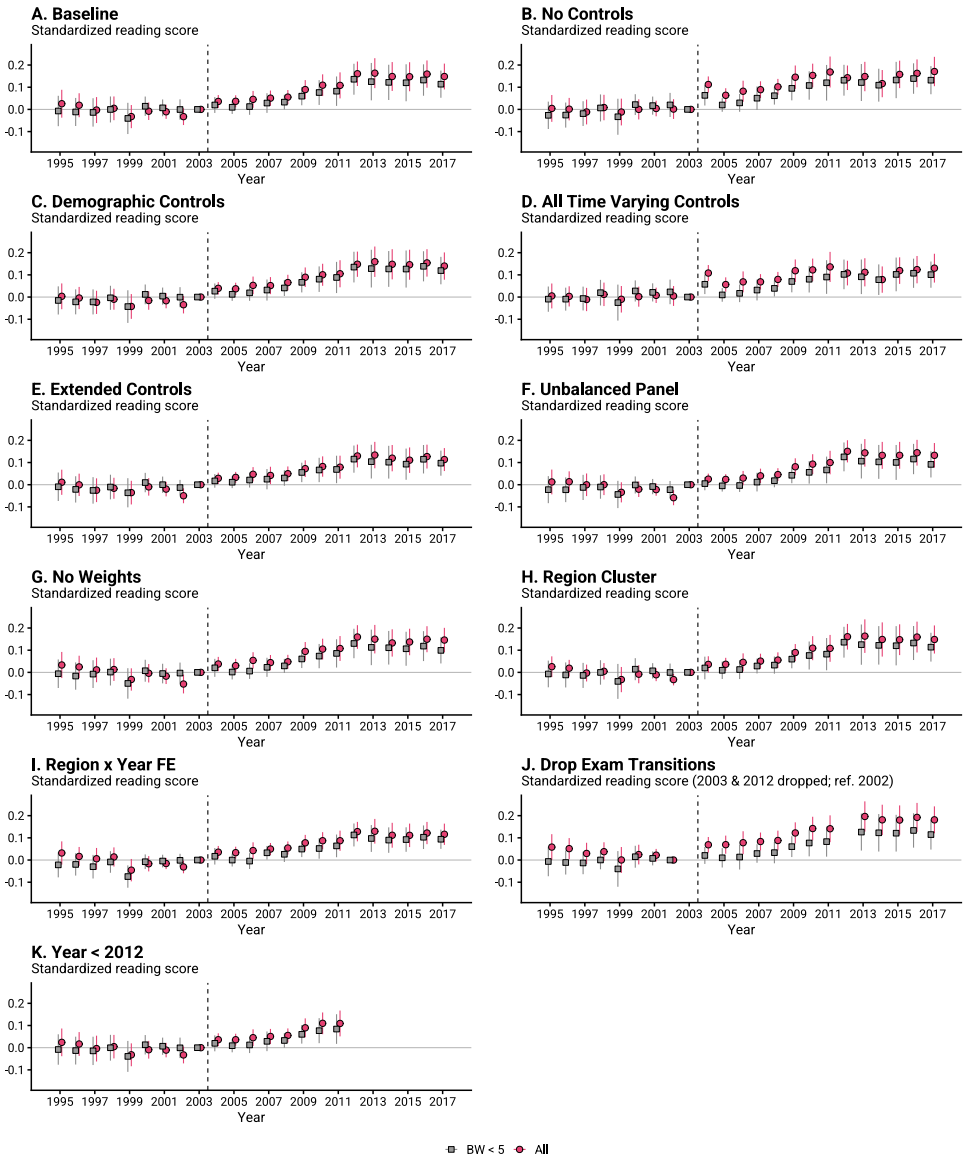
Notes: Each panel reports 2SLS estimates of the effect of $20 million of cumulative RF funding (2025 dollars) on the indicated outcome: test scores by grade, high school outcomes, college outcomes, labor force participation, and wages by age. Circles use the full sample; squares restrict to districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A15: Effect of Reading First eligibility on math decile membership and transitions, grades 3–8



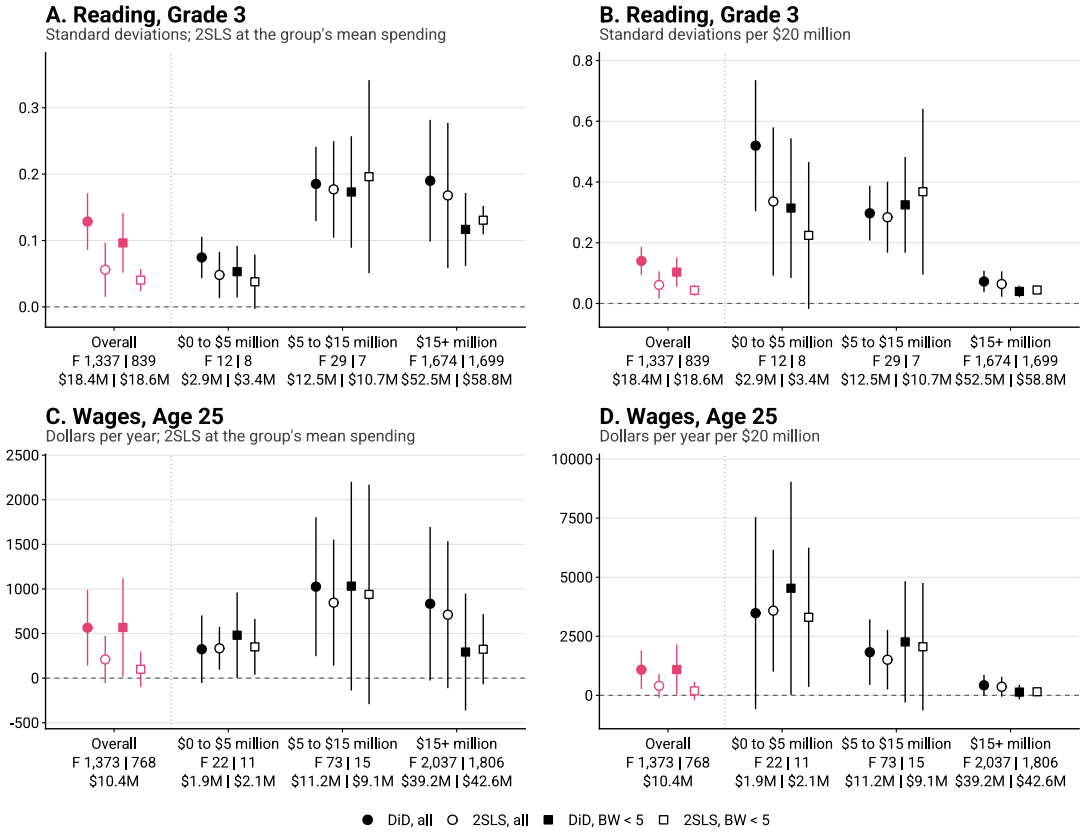
Notes: Panels A and B report difference-in-differences estimates of the effect of RF eligibility on grade-3 and grade-8 math decile membership (percentage points); circles use the full sample and squares restrict to districts within five percentage points of the 13.6% fail rate threshold. Panel C reports the effect on each grade-3-to-grade-8 decile transition (row = grade-3 decile, column = grade-8 decile). Standard errors are clustered at the district level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A16: Event studies under every robustness specification of Figure 2



Notes: Each panel re-estimates the event study of Panel C in Figure 2, for each of the DiD robustness specifications illustrated in Panel D of Figure 2. Panel A of this figure is identical to Panel C of Figure 2, and illustrates our baseline specification for reference. Every panel shows both the full sample (circle) and districts within five percentage points of the 13.6% fail rate threshold (squares). Bars represent 95% confidence intervals based on standard errors clustered at the district level. The last row of Panel D Figure 2 is illustrated as it's own event study in Figure A8.

Figure A17: Treatment effects by three groups of cumulative predicted Reading First dollars, estimated at the group's mean spending and per $20 million



Notes: Each panel reports difference-in-differences (filled markers) and 2SLS (hollow markers) estimates of the effect of Reading First on the indicated outcome, separately for eligible districts in three groups of cumulative predicted RF dollars (2025 dollars): $0 to $5 million (338 districts), $5 to $15 million (16 districts), and $15 million or more (5 districts). Estimates labeled "Overall" in red, pool all eligible districts. Each coefficient comes from a separate regression on untreated students plus the group's eligible students. Panels A and C evaluate the 2SLS slope at the group's mean realized dollars, reported below each group with the first-stage F statistic (full sample | BW < 5), so both markers are level effects. Panels B and D rescale every estimate to $20 million. Wage panels use third-grade cohorts through 2014, and the sample matches Appendix Figure A14. Circles use the full sample; squares restrict to districts within five percentage points of the 13.6% fail rate threshold. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Table B1: Pre-period descriptive statistics: campus and district characteristics

	Full sample (1)	Bandwidth < 5 (2)
A. Campus characteristics (grade 3, 1995–2002 mean)		
Standardized reading score	0.010	-0.001
	(0.341)	(0.299)
Standardized math score	0.012	-0.002
	(0.345)	(0.304)
Enrollment (grade 3)	88	90
	(44)	(39)
Share Black	0.132	0.147
	(0.178)	(0.183)
Share Hispanic	0.367	0.394
	(0.301)	(0.290)
Share White	0.398	0.351
	(0.302)	(0.274)
Share Asian/Native American	0.024	0.024
	(0.039)	(0.039)
Share economically disadvantaged	0.556	0.585
	(0.287)	(0.269)
Share Title I	0.598	0.613
	(0.392)	(0.395)
Share ESL	0.045	0.042
	(0.069)	(0.064)
Share special education	0.120	0.119
	(0.036)	(0.036)
Share gifted/talented	0.077	0.074
	(0.067)	(0.074)
2002 third-grade fail rate (%)	13.4	13.2
	(9.7)	(8.1)
B. District finance characteristics (1995–2002 mean)		
Total revenue per K–3 student ($)	$38,494	$36,588
	(21316)	(9754)
K–3 enrollment (district)	1,004	1,659
	(3538)	(4769)
Campuses	3,047	1,707
Districts	1,128	415

Notes: Sample is the balanced campus panel, grade 3, 1995–2002 (pre-Reading First). Columns report enrollment-weighted pre-period means over all campuses, eligible and ineligible together, with weighted standard deviations in parentheses. Column (2) restricts to campuses within bandwidth < 5 of the eligibility cutoff. Panel B uses each district's 1995–2002 mean (one observation per district) and is weighted by K–3 enrollment. Dollar figures are in 2025 dollars.

Table B2: Effect of Reading First eligibility on non-Reading First school district revenue

	DiD		DiD, separate periods	
	(1)	(2)	(3)	(4)
A. Non-reading first revenue per student				
1(RF Eligible) × 1(year>2003)	223.97	196.74		
	(178.82)	(223.85)		
1(RF Eligible) × 1(2004–2011)			155.08	221.07
			(174.11)	(195.13)
1(RF Eligible) × 1(2012 and after)			310.83	166.22
			(205.89)	(283.72)
Observations	24,907	8,630	24,907	8,630
District FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Weights	All students	All students	All students	All students

Notes: This table tests whether RF eligibility affected non-RF-specific revenue. Each column reports a separate regression of non-RF district revenue per total student (current-year enrollment) on RF eligibility measures. Columns (1)–(2): pooled difference-in-differences. Columns (3)–(4): separate treatment effects for the during-RF (2004–2011) and post-RF (2012 and after) periods. Odd columns use all districts; even columns restrict to districts within five percentage points of the 13.6% fail rate. All specifications include district and year fixed effects, weighted by total district enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: Effect of Reading First eligibility on third-grade reading, all test takers

	DiD		DiD, sep. periods		Endogenous		Reduced form		2SLS	
	All	BW < 5	All	BW < 5	All	BW < 5	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
RF Post	0.097***	0.071***								
	(0.023)	(0.027)								
RF Eligible × During RF			0.056***	0.036						
			(0.020)	(0.023)						
RF Eligible × After RF			0.155***	0.121***						
			(0.031)	(0.038)						
Cumulative RF $ per $20 million (Endogenous)					0.055***	0.038***				
					(0.018)	(0.007)				
Cumulative Predicted RF $ per $20 million							0.078***	0.057***		
							(0.026)	(0.009)		
Cumulative RF $ per $20 million (2SLS)									0.055***	0.041***
									(0.019)	(0.007)
First-stage F-stat									1,775	993
Observations	59,569	33,465	59,569	33,465	59,569	33,465	59,569	33,465	59,569	33,465
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Sample: the balanced campus panel (campuses observed in every year 1995–2017), the paper's common school sample. The outcome is standardized third-grade reading among ALL grade-3 test takers that year, including retakes and repeat-grade attempts. This differs from the main table that restricts to first-time test takers and their first tests. Within each specification, the first column uses the full sample and the second restricts to districts with predicted-funding bandwidth < 5. DiD reports the pooled DiD; DiD, sep. periods separates the during-RF (2004–2011) and post-RF (2012 and after) periods; Endogenous reports OLS on actual cumulative RF $; Reduced form reports the reduced form on predicted cumulative RF $; 2SLS instruments actual with predicted cumulative RF $ (dollars in $ millions throughout; the Endogenous, Reduced form, and 2SLS coefficients are per $20 million). All models include campus and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects (the paper's baseline controls). District-clustered SEs in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4: First-stage estimates across robustness specifications

First-stage IV estimates		
	All	Bandwidth < 5
	(1)	(2)
A. Baseline		
Coefficient	1.419***	1.403***
	(0.034)	(0.048)
F-stat (instrument)	1711.2	863.1
Observations	60,619	34,240
B. No Controls		
Coefficient	1.410***	1.397***
	(0.038)	(0.048)
F-stat (instrument)	1382.9	829.9
Observations	60,619	34,240
C. Demographic Controls		
Coefficient	1.420***	1.403***
	(0.037)	(0.047)
F-stat (instrument)	1485.0	908.1
Observations	60,619	34,240
D. All Time Varying Controls		
Coefficient	1.413***	1.399***
	(0.036)	(0.047)
F-stat (instrument)	1526.3	887.6
Observations	60,619	34,240
E. Extended Controls		
Coefficient	1.414***	1.401***
	(0.042)	(0.038)
F-stat (instrument)	1159.6	1392.2
Observations	60,619	34,240
F. Unbalanced Panel		
Coefficient	1.416***	1.398***
	(0.036)	(0.047)
F-stat (instrument)	1537.9	890.0
Observations	79,898	43,266
G. No Weights		
Coefficient	1.416***	1.397***
	(0.039)	(0.051)
F-stat (instrument)	1319.5	755.1
Observations	60,619	34,240
H. Region Cluster		
Coefficient	1.419***	1.403***
	(0.034)	(0.044)
F-stat (instrument)	1758.6	1015.6
Observations	60,619	34,240
I. Region x Year FE		
Coefficient	1.418***	1.386***
	(0.027)	(0.025)
F-stat (instrument)	2658.3	3111.7
Observations	60,619	34,240
J. Drop Exam Transitions		
Coefficient	1.416***	1.399***
	(0.035)	(0.048)
F-stat (instrument)	1654.6	858.7
Observations	55,347	31,262
K. Year < 2012		
Coefficient	1.365***	1.346***
	(0.033)	(0.047)
F-stat (instrument)	1685.0	834.9
Observations	44,803	25,306

Notes: Each panel reports the first-stage coefficient from regressing cumulative RF expenditures on cumulative predicted RF dollars, for each robustness specification shown in the main reading figure's IV robustness panel. Column (1) uses all districts; column (2) restricts to districts within five percentage points of the 13.6% fail-rate eligibility cutoff. Standard errors clustered at the district level in parentheses. Instrument F-statistics (Wald) are reported below each coefficient. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B5: Effect of Reading First eligibility on third-grade math

	Full sample	Bandwidth < 5
	(1)	(2)
A. DiD		
1(RF Eligible) × 1(year>2003)	0.109***	0.084***
	(0.028)	(0.032)
B. DiD, separate periods		
1(RF Eligible) × 1(2004–2011)	0.067***	0.040
	(0.024)	(0.025)
1(RF Eligible) × 1(2012 and after)	0.163***	0.140***
	(0.036)	(0.045)
C. 2SLS, cumulative RF $		
Cumulative RF $ per $20 million (2SLS)	0.069***	0.056***
	(0.017)	(0.008)
First-stage F-stat	1,574	848
Observations	60,619	34,240
Campus FE	Yes	Yes
Year FE	Yes	Yes

Notes: The outcome is standardized third-grade math among first-time test takers. The sample is the balanced campus panel (campuses observed in every year 1995–2017), the paper's common school sample. Column (2) restricts to districts with predicted-funding bandwidth < 5. Panels report the pooled DiD, the split-period DiD, and 2SLS instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; the 2SLS coefficient is per $20 million. All models include campus and year fixed effects, the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects (the paper's baseline controls), and are weighted by the number of students taking the math exam. District-clustered SEs in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B6: Evaluating the effect of Reading First when controling for non-Reading first district revenue (robustness check)

	Table 2 1995–2017	No control 1997–2017	Non-RF total 1997–2017	Non-RF federal 1997–2017
	(1)	(2)	(3)	(4)
A. DiD				
1(RF Eligible) × 1(year>2003)	0.110***	0.117***	0.117***	0.116***
	(0.022)	(0.022)	(0.022)	(0.021)
Non-RF revenue control			0.000	0.003*
			(0.000)	(0.002)
B. 2SLS (cumulative RF $ per $20 million)				
Cumulative RF $	0.059***	0.061***	0.061***	0.060***
	(0.019)	(0.020)	(0.020)	(0.020)
Non-RF revenue control			0.000	0.004**
			(0.000)	(0.002)
First-stage F	1,711.2	1,706.4	1,412.0	1,622.8
Observations	60,619	55,350	55,350	55,350
Campus FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Notes: The outcome is the standardized third-grade reading score among first-time test takers. Column (1) is the Table 2 specification on the full 1995–2017 panel. Column (2) repeats it on the 1997–2017 window over which district finance data exist; column (3) adds non-Reading First TOTAL district revenue as a control and column (4) non-Reading First FEDERAL district revenue, both on that window. The gap between columns (1) and (2) is the window, not the controls: the two dropped pre-years carry small positive eligible-minus-ineligible differences in the event study, so removing them raises the post-versus-pre contrast. Controls are per K–3 student, in thousands of real 2025 dollars. Panel A is the pooled eligibility DiD; Panel B instruments cumulative Reading First dollars with cumulative predicted Reading First dollars, both in $ millions; the cumulative RF $ coefficient and its standard error are reported per $20 million. All models are estimated on the balanced campus panel and include campus and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects (the paper's baseline controls), and are weighted by the number of reading test takers. District-clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B7: Effect of Reading First eligibility on Student Success Initiative margins

	All (1)	Bandwidth < 5 (2)
Any positive Fund 404 amount	0.0259	0.0174
	(0.0199)	(0.0180)
Observations	12,375	4,576
Net Fund 404 $ per 100 fixed 2002 readers	$17,825	-$2,497
	($11,314)	($17,001)
Observations	12,375	4,576

Notes: Each cell is the coefficient on 1(RF eligible) $\times$ 1(year $\geq$ 2003) from a separate regression; 2003 dates the mechanisms to the Student Success Initiative's onset, one year before Reading First outcomes. Fund 404 is TEA's dedicated SSI fund. Both rows use a district-year panel over 2001–2011 with district and year fixed effects, weighted by the district's fixed 2002 count of grade-3 reading takers (also the per-100-reader denominator, so the dollar outcome cannot move through enrollment); district-years without a Fund 404 accounting row are zeros, which the construction verifies is a true absence rather than missing data. Standard errors clustered by district in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B8: Effect of Reading First eligibility on third-grade reading by campus $PM_{2.5}$ concentration, 2000-2016

	Above median $PM_{2.5}$		Below median $PM_{2.5}$	
	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)
A. DiD				
1(RF Eligible) × 1(year>2003)	0.115***	0.086***	0.107***	0.080***
	(0.032)	(0.032)	(0.024)	(0.030)
B. DiD, separate periods				
1(RF Eligible) × 1(2004–2011)	0.082***	0.054**	0.063***	0.037
	(0.026)	(0.021)	(0.020)	(0.024)
1(RF Eligible) × 1(2012 and after)	0.156***	0.126**	0.164***	0.135***
	(0.044)	(0.048)	(0.033)	(0.044)
C. 2SLS, cumulative RF $				
Cumulative RF $ per $20 million (2SLS)	0.051***	0.038***	0.152***	0.349*
	(0.016)	(0.006)	(0.042)	(0.199)
First-stage F-stat	4,003	2,383	103	6
Observations	30,311	18,535	30,308	15,705
Campus FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Notes: Sample: the balanced campus panel (campuses observed in every year 1995-2017), the paper's common school sample, but this time constrained to the 2000-2016 period due to the limitation in availability of $PM_{2.5}$ satellite data (Di et al., 2016). The outcome is standardized third-grade reading among first-time test takers, in the main reading table's panel layout. Campuses are split at the median of their own estimation-sample annual-average $PM_{2.5}$ (each campus's mean $PM_{2.5}$, 2000-2016, median taken across campuses in the reading estimation sample); campuses at or above the median are "Above," the rest "Below." Within each pollution group, the first column uses the full sample and the second restricts to districts with predicted-funding bandwidth < 5. Panels report the pooled DiD, the split-period DiD, and 2SLS instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; the 2SLS coefficient is per $20 million. All models include campus and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects (the paper's baseline controls), and are weighted by the reading-tested count (n_read). District-clustered SEs in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B9: Effect of Reading First eligibility on Reading First school district broad expenditure categories

	All tracked expenditures		Salary, wages & benefits		Travel		Prof. purchased services		Supplies & materials		Misc. operating	
	All	BW < 5	All	BW < 5	All	BW < 5	All	BW < 5	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. DiD												
1(RF Eligible) × 1(year>2003)	74.35***	58.87***	41.12***	33.87***	1.50***	0.99***	4.79***	2.95***	26.41***	22.06***	0.55	-1.00
	(8.44)	(11.40)	(4.30)	(5.93)	(0.36)	(0.35)	(1.05)	(0.70)	(3.57)	(4.61)	(1.72)	(2.43)
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
B. DiD, separate periods												
1(RF Eligible) × 1(2004–2011)	131.47***	105.38***	72.50***	60.28***	2.68***	1.81***	8.35***	5.12***	46.78***	39.49***	1.16	-1.31
	(14.64)	(20.04)	(7.48)	(10.52)	(0.62)	(0.62)	(1.81)	(1.25)	(6.23)	(8.17)	(2.73)	(3.86)
1(RF Eligible) × 1(2012 and after)	1.68	0.23	1.18	0.57	-0.01	-0.03	0.25	0.21	0.48	0.08	-0.22	-0.60
	(1.47)	(1.69)	(0.78)	(0.78)	(0.05)	(0.04)	(0.17)	(0.18)	(0.53)	(0.60)	(0.45)	(0.66)
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
C. 2SLS												
Cumulative RF $ per $20 million (2SLS)	2.13***	1.76**	2.90**	1.95***	0.01	-0.11**	-0.96	-0.12***	0.48	0.46	-0.31	-0.43
	(0.75)	(0.75)	(1.17)	(0.53)	(0.14)	(0.05)	(0.82)	(0.04)	(0.82)	(0.41)	(0.20)	(0.31)
First-stage F-stat	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weights	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students

Notes: Each pair of columns reports a separate regression of RF-specific expenditures per K–3 student. Columns (1)-(2) report results across all tracked expenditure categories. The remaining pairs of columns break up the all tracked expenditures category into its five individual components, namely: salary, wages & benefits (columns 3-4), travel (columns 5-6), professional purchased services (columns 7-8), supplies and materials (columns 9-10), and miscellaneous operating expenditures (columns 11-12). Panel A shows pooled difference-in-differences estimates. Panel B estimates separate treatment effects for the during-RF (2004–2011) and post-RF (2012 and after) periods. Panel C reports 2SLS estimates instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; coefficients are per $20 million of cumulative RF $, with the first-stage F-statistic reported below the coefficient. Within each outcome, the first column uses the full sample and the second restricts to districts within five percentage points of the 13.6% fail rate threshold. All specifications include district and year fixed effects and are weighted by K–3 enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B10: Effect of Reading First eligibility on Reading First school district supplies and materials related expenditure categories

	Supplies & materials		Books		General supplies		Testing materials	
	All	BW < 5	All	BW < 5	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. DiD								
1(RF Eligible) × 1(year>2003)	26.41***	22.06***	13.51***	11.02***	12.65***	10.85***	0.23***	0.19**
	(3.57)	(4.61)	(2.61)	(3.49)	(2.63)	(3.65)	(0.07)	(0.10)
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
B. DiD, separate periods								
1(RF Eligible) × 1(2004–2011)	46.78***	39.49***	23.90***	19.51***	22.43***	19.63***	0.41***	0.35**
	(6.23)	(8.17)	(4.55)	(6.12)	(4.59)	(6.50)	(0.12)	(0.17)
1(RF Eligible) × 1(2012 and after)	0.48	0.08	0.30	0.32	0.21	-0.22	-0.01	-0.01
	(0.53)	(0.60)	(0.26)	(0.32)	(0.48)	(0.50)	(0.01)	(0.01)
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
C. 2SLS								
Cumulative RF $ per $20 million (2SLS)	0.48	0.46	0.36	0.44*	0.08	-0.01	0.04	0.03
	(0.82)	(0.41)	(0.75)	(0.25)	(0.13)	(0.22)	(0.04)	(0.05)
First-stage F-stat	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weights	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students	K–3 students

Notes: Each pair of columns reports a separate regression of RF-specific expenditures per K–3 student by specific supply subcategory. Panel A: pooled difference-in-differences. Panel B: separate treatment effects for the during-RF (2004–2011) and post-RF (2012 and after) periods. Panel C: 2SLS estimates instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; coefficients are per $20 million. Within each outcome, the first column uses the full sample and the second restricts to districts with predicted-funding bandwidth < 5. All specifications include district and year fixed effects, weighted by K–3 enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B11: Effect of Reading First eligibility on Reading First school district book-related expenditure categories

	Books		Reading materials		Library books		Textbooks	
	All	BW < 5	All	BW < 5	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. DiD								
1(RF Eligible) × 1(year>2003)	13.51***	11.02***	12.64***	10.68***	0.17**	0.15	0.70*	0.19
	(2.61)	(3.49)	(2.66)	(3.53)	(0.08)	(0.09)	(0.38)	(0.41)
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
B. DiD, separate periods								
1(RF Eligible) × 1(2004–2011)	23.90***	19.51***	22.25***	18.80***	0.31**	0.27*	1.34**	0.44
	(4.55)	(6.12)	(4.66)	(6.20)	(0.14)	(0.16)	(0.66)	(0.69)
1(RF Eligible) × 1(2012 and after)	0.30	0.32	0.41*	0.45	-0.01	-0.01	-0.11	-0.13
	(0.26)	(0.32)	(0.24)	(0.27)	(0.01)	(0.01)	(0.08)	(0.13)
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
C. 2SLS, cumulative RF $								
Cumulative RF $ per $20 million (2SLS)	0.36	0.44*	0.41	0.46*	-0.02**	-0.01*	-0.03	-0.01
	(0.75)	(0.25)	(0.73)	(0.26)	(0.01)	(0.00)	(0.05)	(0.03)
First-stage F-stat	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48	1687.92	1158.48
Observations	23,831	8,628	23,831	8,628	23,831	8,628	23,831	8,628
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weights	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students

Notes: Each pair of columns reports a separate regression of RF-specific expenditures per K–3 student. Columns (1)-(2) report results across the books category. The remaining pairs of columns break up the books category into its three individual components, namely: reading materials (columns 3-4), library books (columns 5-6), and textbooks (columns 7-8). Panel A shows pooled difference-in-differences estimates. Panel B estimates separate treatment effects for the during-RF (2004–2011) and post-RF (2012 and after) periods. Panel C reports 2SLS estimates instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; coefficients are per $20 million. The first-stage F-statistic is reported below the coefficient. Within each outcome, the first column uses the full sample and the second restricts to districts within five percentage points of the 13.6% fail rate threshold. All specifications include district and year fixed effects and are weighted by K–3 enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B12: Long-run effects on high school completion, college, and age-25 labor-market outcomes

	High School		College Enrollment			College	Age 25	
	On-Time Diploma	Any Completion	Any	2yr	4yr	BA Receipt	Employed	Earnings ($)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. DiD (FEOLS)								
1(RF Eligible) × 1(3rd grade Year ≥ 2004)	3.09***	2.52***	2.50**	-0.15	1.70*	2.97***	-0.14	$554**
	(0.41)	(0.39)	(1.02)	(1.17)	(1.00)	(0.64)	(0.19)	($218)
B. 2SLS: Cumulative RF $ (per $20 million)								
Cumulative RF $ per $20 million (2SLS)	1.25***	1.08***	2.16***	0.76*	2.06***	1.94***	-0.50***	$366
	(0.23)	(0.13)	(0.26)	(0.42)	(0.45)	(0.23)	(0.12)	($251)
First-stage F-stat	1,480	1,480	1,493	1,493	1,493	1,493	1,407	1,407
C. Poisson (Earnings Only)								
Poisson coefficient (log points)	–	–	–	–	–	–	–	0.022***
								(0.007)
Implied $ effect at mean	–	–	–	–	–	–	–	$717
Observations	5,555,625	5,557,260	6,810,303	6,810,303	6,810,303	6,810,303	3,693,898	3,694,781
Pre-RF Cohort Mean	88.2	91.8	57.6	44.3	45.6	24.3	78.3	$32,500
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Third-Grade-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Unit of observation is the student, in third-grade cohorts 1994–2014 (cohorts after 2014 cannot yet appear in the HS/college files); grade-3 repeaters enter once, at their first attempt. Columns (1)-(2) report effects on on-time standard HS diploma and any HS completion respectively, Columns (3)-(5) report estimates for ever enrolled in college (any/2-year/4-year), ever enrolled in a 2-year college, ever enrolled in a 4-year college respectively. Column (6) examines the effect on BA receipt, Columns (7)-(8) report the effects on whether the student was employed (observed UI wages) at age 25 and their UI-covered earnings at age 25 (2025 dollars) respectively. Both age-25 outcomes are defined only for students attached to the Texas UI universe (a wage record at some age 18–25) whose age-25 year falls inside the UI window (2010–2025); within that group, unobserved age-25 earnings are imputed as zero and employment as not employed. Panel A shows pooled difference-in-differences estimates, for cohorts treated if their 3rd-grade year is ≥ 2004. Panel B shows 2SLS estimates, instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; coefficients are per $20 million. Panel C reports a Poisson (pseudo-maximum-likelihood) version of the earnings regression; the implied $ effect evaluates the proportionate Poisson coefficient at the pre-RF mean. All models include campus and third-grade-year fixed effects, with district-clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B13: College enrollment by in state vs. out of state institution, and inclusion in the Texas wage records

		First Institution				Texas
	Any College	In Texas	Out of State	Unidentified	Out of State, 4yr	Ever in TX UI
	(1)	(2)	(3)	(4)	(5)	(6)
A. DiD (FEOLS)						
1(RF Eligible) × 1(3rd-Grade Year ≥ 2004)	2.50**	2.02*	0.15**	0.33	0.11	2.08*
	(1.02)	(1.20)	(0.06)	(1.12)	(0.12)	(1.11)
B. 2SLS: Cumulative RF $ (per $20 million)						
Cumulative RF $ per $20 million (2SLS)	2.16***	1.74***	0.17***	0.25	0.24***	2.36***
	(0.26)	(0.37)	(0.04)	(0.50)	(0.05)	(0.35)
First-stage F-stat	1,493	1,493	1,493	1,493	1,493	1,493
Observations	6,810,303	6,810,303	6,810,303	6,810,303	6,810,303	6,810,303
Pre-RF Cohort Mean	57.6	42.1	2.3	13.2	3.0	78.7
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes
Third-Grade-Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Unit of observation is the student, in third-grade cohorts 1994–2014, with grade-3 repeaters entering once, at their first attempt; the sample, filters and specification are identical to the long-run outcomes table. All outcomes are indicators and all coefficients are reported in percentage points. Column (1) is enrollment in any college observed in the data. Columns (2)–(4) split that enrollment by the location of the FIRST institution attended, using an IPEDS crosswalk from FICE code to state: in Texas, outside Texas, or institution not identifiable (the enrollment record carries no usable FICE code, or the code is absent from the crosswalk). Columns (2)–(4) sum to column (1) by construction. Institutions operating in Texas and in other states under a single FICE code are counted as in Texas, which understates out-of-state enrollment. Column (5) restricts to first enrollment at a four-year institution outside Texas. Column (6) indicates that the student is ever observed in Texas unemployment-insurance wage records at any age 18–25; students who leave Texas permanently never appear, so a negative coefficient is the earnings-side counterpart of out-of-state migration. Enrollment is observed through Texas Higher Education Coordinating Board records and National Student Clearinghouse files, the latter covering enrollment at institutions outside Texas for 2009–2017; enrollment out of state and outside that window is not observed, so column (3) is a lower bound. Panel A is difference-in-differences, with a cohort treated once its third-grade year is ≥ 2004. Panel B is 2SLS instrumenting cumulative RF $ with cumulative predicted RF $, both in $ millions; coefficients are per $20 million. All models include campus and third-grade-year fixed effects, with district-clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B14: Age-25 wages and expected-wage channels

	Actual	Expected wage at 25, given:			Complex
	wage at 25	College (FICE)	Field (CIP)	Industry (NAICS)	prediction
	(1)	(2)	(3)	(4)	(5)
A. Effect of RF on actual and expected wages at 25					
Reading First (post)	554.1**	646.4***	507.2***	294.0***	663.9***
	(217.8)	(119.7)	(89.2)	(92.8)	(146.8)
B. Actual wage at 25, controlling for each expected-wage channel					
Reading First (post)	554.1**	-54.1	69.3	264.8	-31.7
	(217.8)	(211.3)	(195.9)	(173.5)	(165.4)
Expected-wage control	–	0.941***	0.956***	0.984***	0.882***
		(0.007)	(0.008)	(0.010)	(0.006)
Observations	3,694,781	3,694,721	3,694,780	3,694,781	3,694,781
Campus FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: Individual-level regressions on the age-25 wage panel: first-time third-grade test takers matched to state UI wage records at age 25. The actual wage (column 1) is the age-25 wage, imputed to zero only when the individual has at least one observed wage and age 25 falls within UI coverage. Columns (2)–(4) replace the outcome with the expected wage given the individual's college attended (FICE), field of study (CIP), or industry at 25 (NAICS): the untreated-sample mean age-25 wage within that cell, with explicit categories for non-attendance and non-employment. Column (5) uses the complex predicted wage from a Poisson fixed-effects model combining all channels. Panel A reports the effect of RF eligibility × post-2004 on each wage measure. Panel B's outcome is the actual age-25 wage in every column, adding the column's expected-wage measure as a control; column (1) of Panel B is the unadjusted baseline (its "control" would be the outcome itself), indicated by –. All models include campus and third-grade-year fixed effects. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C Data and Estimation Details

C.1 Tax Elections

We take the record of maintenance and operations (M&O) tax rate elections from the replication files of Abott et al. (2020). These data cover Texas tax ratification and bond measures from 2000 to 2015. We keep the tax measures, match districts to TEA county-district numbers, and record the year of the first successful tax election for each district. Bond elections are not part of the treatment.

Given that election records from Abott et al. (2020) end in 2015, we extend the sample through 2019 by identifying later elections from the M&O tax rates that districts adopt each year from TEA publications. We are able to identify a successful election by noting that the adoption rate may only exceed the rollback ceiling once voters have approved it. We date a passage to the first year we observe a district's M&O rate crossing the $1.04 rollback ceiling, having been at or below it the year before.

We can also identify M&O tax rate increases by assessing whether the district's M&O rate rises by at least $0.11 net of the statewide change in that year. We obtain the $0.11 threshold by regressing the year-over-year change in a district's M&O rate on a passage indicator and year fixed effects, over the years the Abott et al. (2020) records cover. When applying this threshold back to the districts with observed elections, we can identify the correct passage year for 91 percent of districts (94 percent within a year). [16]

We use the Abott et al. (2020) records as being authoritative for the years 2000-2015 and the estimated rate-detected year for 2016–2019. Thus, treatment timing is a hybrid of both sources. In all, 580 districts are treated, 410 from the records and 170 from rates. The average treated district raised its M&O rate by $0.103 while cutting its debt-service rate by $0.059. Because Texas matches M&O tax effort with state aid but not debt-service taxes, marginal changes between the two rates raise state revenue even when the total tax rate changes little. This is one reason the total revenue effect in Panel A of Table C1 exceeds the local-revenue effect alone.

C.1.1 Finance outcomes

Revenue and expenditure data come from the Public Education Information Management System (PEIMS) financial records described in Section 3.3, deflated to 2025 dollars with the CPI-U. PEIMS contains object codes that identify its source and fund, and function codes describing where the money is held and how it is to be used. We classify revenue as local, state, or federal using the object code independent of fund or function. [17] Salary, wages and

[16] Nearly all of the remaining misses date to 2007. Rates had been compressed from the $1.50 cap the year before, so a district adopting the new $1.17 maximum neither crosses $1.04 nor clears $0.11 once the statewide compression is netted out.

[17] Money that reaches districts through county-level entities (intermediate sources) is negligible in Texas, and we include it in local revenue.

benefits, travel, professional and purchased services, supplies and materials, and miscellaneous operating follow the same object-code groupings used for Reading First expenditure. General supplies, testing materials, and books are components of supplies and materials, and books is the sum of library books, textbooks, and reading materials.

The results in Table C1 are estimated at the district-year level. Each expenditure record in PEIMS carries an organization code that assigns it either to a campus or to a district-level unit such as central administration. Panel A sums the revenue records assigned to the district as a whole. Panel B sums the expenditure records assigned to campuses, so all tracked expenditure is the district's total campus-assigned spending in the five listed categories rather than total district spending. Note that we cannot observe where campus expenditures are allocated to, only that expenditures were allocated to campuses. Each outcome is divided by the district's total enrollment in that year, from the TEA enrollment files, and the same count serves as the regression weight.

C.1.2 Estimation

We estimate the interaction-weighted estimator of Sun and Abraham (2021) on the district-year panel, with district and year fixed effects, enrollment weights, and standard errors clustered by district. A district's cohort is the passage year of its first successful tax election. Districts with no election are the never-treated comparison group, and not-yet-treated districts contribute in the usual way. Table C1 reports the overall average treatment effect on the treated. Panel A of Figure C1 plots the same estimator by event time, with the year before passage as the reference period. Panel B is a two-way fixed-effects event study of federal revenue per student resulting from Reading First, estimated among eligible districts with all other districts as the reference group.

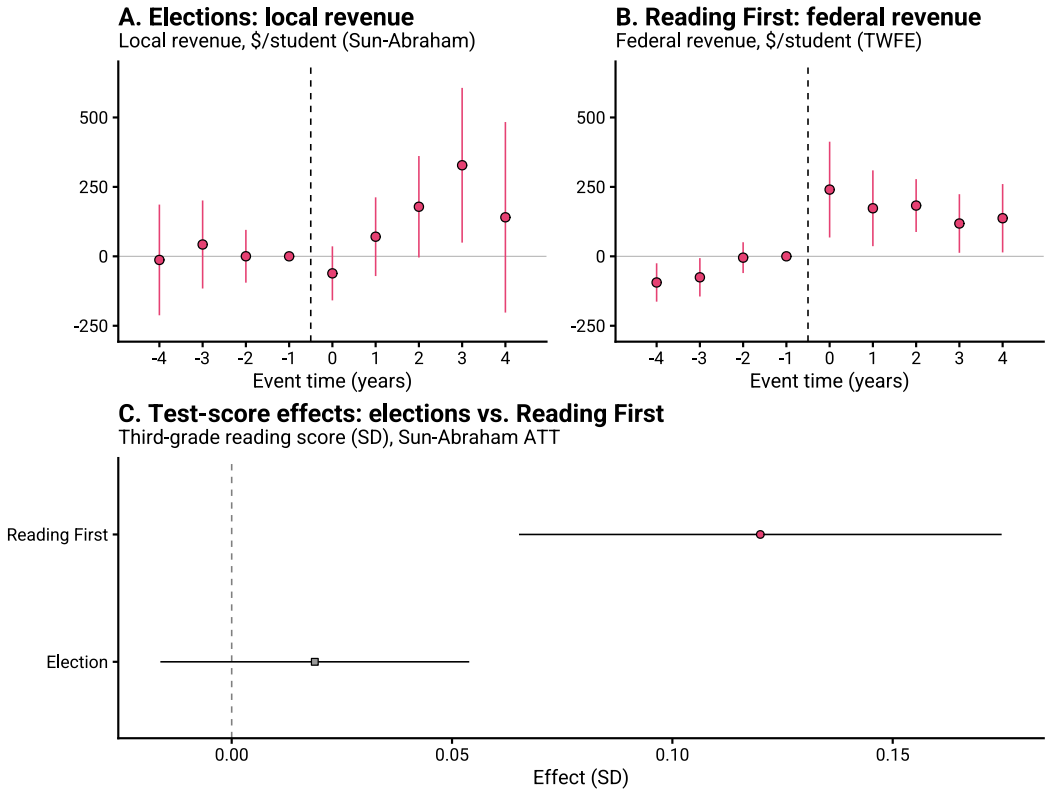
Panel C compares the two treatments on third-grade reading. Both estimates come from the same campus-year panel for 2000–2019 with campus and year fixed effects, weighted by the number of tested students and clustered by district. Passing an election changes reading scores by 0.019 standard deviations, while Reading First changes them by 0.120.

Table C1: The effect of passing a maintenance and operations tax election on district revenue and expenditure, 2000–2019

	Tax election
A. Revenue ($ per student)	
Total revenue	427.7***
	(125.8)
Local revenue	320.9*
	(164.6)
B. Expenditure ($ per student)	
All tracked expenditure	339.6***
	(130.4)
Salary, wages & benefits	189.5***
	(53.0)
Travel	7.0***
	(2.2)
Prof. purchased services	125.5
	(111.2)
Supplies & materials	24.7
	(27.2)
General supplies	16.4
	(24.1)
Testing materials	−2.0***
	(0.7)
Books (total)	−4.2
	(5.5)
Library books	0.3
	(1.7)
Textbooks	−3.1
	(4.4)
Reading materials	−1.5
	(3.9)
Misc. operating	−7.1
	(13.9)
Observations	24,112
Districts	1,343
Passed a tax election	580
District FE	Yes
Year FE	Yes
Weights	Enrollment

Notes: District-year panel, 2000–2019. Sun-Abraham overall ATT treating a district in the passage year of its first successful maintenance and operation tax election. Passage years are taken from the AKLP election records where those exist (2000–2015) and otherwise detected from the district's adopted tax rates for 2016–2019: the first year the M&O rate crosses the $1.04 rollback ceiling or rises by at least $0.11 relative to the statewide change, a rule that returns the recorded year for 91 percent of districts with an AKLP record. Districts that have not yet passed or never passed an election are controls. Outcomes are district revenue and school expenditure by category, per enrolled student; general supplies, testing materials, and books are components of supplies and materials, and books is the sum of library books, textbooks, and reading materials. All models include district and year fixed effects, weight by district enrollment, and cluster standard errors by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure C1: District revenue and test scores: tax elections vs. Reading First



Notes: Panels A and B report event-study estimates of the effect of a passed school district maintenance and operations tax election (Panel A, Sun-Abraham estimator) on local revenue per enrolled student and of Reading First eligibility (Panel B, two-way fixed effects) on federal revenue per enrolled student, 2000–2019. Panel C compares the implied average treatment effects on standardized third-grade reading scores for elections and Reading First. Bars represent 95% confidence intervals based on standard errors clustered at the district level. This figure accompanies Table C1.

C.2 Reading Intensity of Occupations

C.2.1 Occupation scores

Table C2 examines whether Reading First led students to choose more reading-intensive fields of study or industries. In this section, we describe how the reading-intensity scores are built and attached to student records.

The base score is the O*NET importance rating for reading comprehension, on a scale of 1 to 5, from O*NET release 29.2. We collapse O*NET occupation codes to the 763 six-digit occupations of the 2018 Standard Occupational Classification (SOC) by averaging and standardizing the results across those occupations. This is the only standardization in the chain. The field-of-study, college, and industry scores that follow are each averages of these occupation z-scores. These scores are in standard deviations of reading importance across occupations, and their sample means need not be zero.

C.2.2 Field of study, college, and industry scores

We do not observe the occupations students hold. However, we do observe their field of study, the college they attend, and the industry of their employer. Each of these three components is assigned the average reading score of the occupations associated with it. For a field of study, this is the mean occupation score among workers who majored in that field, weighted by how many of them hold each occupation, following Altonji et al. (2016). A nursing degree, for example, receives an O*NET score close to that of registered nurses, while a general business degree receives a blend of the many occupations its graduates enter. The weights are the number of workers in each field-and-occupation pair, which we count in the American Community Survey for 2019 and 2021 through 2024 among employed workers aged 25 to 64 with a bachelor's degree or higher. Fields of degree are mapped to the four-digit Classification of Instructional Programs (CIP) using the Census crosswalk. For the 39 of 441 four-digit CIP codes that the ACS does not cover, we use the unweighted mean over the National Center for Education Statistics CIP-to-SOC crosswalk. The industry score is the employment-weighted mean of the occupation scores within each three-digit NAICS industry, using national employment by occupation and industry from the May 2024 Occupational Employment and Wage Statistics.

Each student's field-of-study score is the score of the four-digit CIP of their first college enrollment. The college score is the mean field-of-study score among the institution's enrollees from never-eligible districts and pre-2004 cohorts, so that it cannot absorb a treatment-induced change in majors. The industry score is the mean, for ages 18 to 25, of the score of the three-digit NAICS industry of the student's main employer in each year, from the Texas Workforce Commission unemployment insurance wage records, where the main employer is the one contributing the majority of a student's income, and missing years are skipped.

C.2.3 Sample and estimation

The sample is the same cohort as in Table B12, students in third grade through 2014, the last cohort for whom we observe age-25 outcomes. We estimate the difference-in-differences specification from Section 4.2 at the student level in Table C2, using campus and third-grade-year fixed effects and standard errors clustered by district. Observation counts differ across outcomes because a student only enters a column if the relevant choice has a score. Even number columns restrict the sample to districts within five percentage points of the eligibility threshold. None of the point estimates in Table C2 are statistically significant, indicating that Reading First did not affect the reading intensity of the choice of CIP, college or employment industry for treated students.

Table C2: Effect of Reading First on the O*NET reading intensity of postsecondary and career choices

	Field of Study (CIP)		College Attended (FICE)		Industry of Employment (NAICS)	
	All	BW < 5	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)	(5)	(6)
1(RF Eligible) × 1(Year ≥ 2004)	-0.004	-0.004	0.000	-0.001	0.002	-0.002
	(0.003)	(0.004)	(0.001)	(0.001)	(0.003)	(0.004)
Observations	2,828,787	1,520,903	3,277,604	1,762,707	5,296,967	2,917,889
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each outcome is the O*NET reading-importance z-score of a post-schooling choice, testing for compositional shifts toward (or away from) reading-intensive paths. Columns (1)–(2): the score of the student's field of study (4-digit CIP) at first college enrollment; sample is enrollees with a scored major. Columns (3)–(4): the average score of the institution (FICE) first attended, computed as the mean field-of-study score among that institution's never-treated and pre-period enrollees; sample is enrollees at a scored institution. Columns (5)–(6): the mean score of the student's 3-digit NAICS industry of main employment across ages 18–25 (UI wage records); sample is workers with a scored industry. Within each outcome, the first column uses the full grade-3 panel (third-grade year ≤ 2014) and the second restricts to districts with fail-rate bandwidth < 5. Each cell is a student-level DiD regression of the outcome on RF eligibility × 1(Year ≥ 2004) with campus and third-grade-year fixed effects; district-clustered SEs in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C.3 Construction of the Marginal Value of Public Funds

To obtain an estimate for the Marginal Value of Public Funds (MVPF) resulting from Reading First, we follow Hendren and Sprung-Keyser (2020) throughout.[18] The MVPF is the ratio of beneficiaries' willingness to pay to the net cost borne by the government,

$$\text{MVPF} = \frac{\text{WTP}}{\text{Net cost}} = \frac{E - \tau E}{1 - \tau E},$$

where E is the present value of the lifetime earnings gain generated per dollar of program spending and τ is the marginal tax-and-transfer rate. Given that the MVPF is a ratio, it is invariant to scale, so every quantity is normalized to one dollar of government spending. The denominator is a *net* cost, where the government pays out a dollar and recovers τE of it via

[18] Their assessment of the MVPF of school spending from Jackson et al. (2016) is particularly useful in our setting.

taxes on the higher future earnings that each dollar produces. When τE reaches one, the net cost is zero, and the MVPF is unbounded, which implies that the policy has paid for itself.

C.3.1 Inputs

The calculation requires five quantities. First is the earnings effect, which we obtain from our difference-in-differences estimate of $554 on earnings at age 25, reported in Table B12, against a mean of $32,500 among students in untreated cells, or a proportional effect of 1.70 percent. Both figures are in 2025 dollars.

Second, we need the program spending per exposed student, which converts our per-student estimate into the per-dollar terms the MVPF requires. Table 1 reports that eligible districts received $132.27 per K–3 student per year while program funds were flowing. A student's exposure can run up to four years from K–3. The cohorts contributing to the age-25 earnings estimate were in third grade between 2004 and 2009, and the earliest cohort had already passed through part of K–3 before funds arrived. Measured directly in the data, these cohorts overlapped with the 2004–2011 funding window for an average of 2.72 years. To be conservative, we round the funding window up to three years. Cumulative spending is therefore approximately $397 per exposed student, in the same 2025 dollars as the earnings estimate.

Third is the marginal tax-and-transfer rate. We use 18.9 percent, obtained by expressing the counterfactual income as a percentage of the federal poverty line and applying the Congressional Budget Office schedule of effective marginal rates, which incorporates federal and state income taxes and the phase-out of transfer programs, and then netting out payroll taxes. This is the same rate their analysis applies to college admissions.

Fourth, the discount rate and real wage growth are 3 percent and 0.5 percent, respectively. Lastly, the age at which spending occurs is taken as 7, the midpoint of K–3 exposure. Benefits are discounted back to this age so that costs and benefits are compared at the same point in time.

C.3.2 Lifetime present value

We now extend the single earnings effect from age 25 through age 65. Given that the wage premium effect of the Reading First treatment at age 25 is 1.70 percent, dividing by the cumulative per-student spending of $397 yields an effect of 0.0043 percent, attributable to a single dollar of spending. We assume that each dollar of program spending was equally productive over time. We then take the mean earnings by age of the 2015 American Community Survey, adjust it to 2025 dollars and to our cohorts' calendar years using 0.5 percent annual real wage growth, and scale the profile so that it matches our own sample's mean at age 25.

When applying the per-dollar earnings effect of 0.0043 percent to counterfactual mean earnings of $32,500 we get $1.40 resulting from a dollar of Reading First by age 25. For all ages, we apply the same percentage. Discounting the gains at each age back to age 25 at 3 percent gives $59.60 per dollar of spending. The money itself was spent when students were

about 7, so we discount another eighteen years to compare the costs and benefits at the same date. This leaves a present value of $35.01 in lifetime earnings per dollar of Reading First spending.

The government recovers part of the $35.01 in lifetime earnings per dollar through taxes and reduced transfers. If we apply an 18.9 percent tax-and-transfer rate, it yields $6.62 of additional revenue per dollar spent on Reading First. Therefore, the net cost is one dollar less $6.62, or −$5.62. Willingness to pay is the earnings gain net of those taxes, $35.01 less $6.62, or $28.39. Because the revenue recovered exceeds the dollar spent, the net cost is negative, and the MVPF is unbounded. Reading First pays for itself.

Table C3 reports how far into adulthood the earnings gains must persist before the program has paid for itself. The first row is the most conservative, where we only consider the age 25 earnings. The earnings gain is $1.40 per dollar of spending, which discounts to $0.82 at the time of exposure. At this age, the government recovers sixteen cents of each dollar, yielding an MVPF of 0.79. Each subsequent row adds one more year of the projected gain, and net cost reaches zero once cumulative recovered revenue equals the dollar spent, which occurs by age 30.

We note that this age 30 threshold is conservative. We hold the proportional earnings effect constant over the life cycle, which existing literature suggests is unlikely. For instance, Mountjoy (2026) estimates the returns to Texas public universities year by year over exactly this range and finds the proportional effect more than doubling between ages 25 and 30. Substituting this trajectory would bring the "pay for itself" threshold forward to age 28. Thus, our constant-effect likely understates the returns to Reading First.

Our estimates are also insensitive to some of the choices made in the MVPF calculation. The program still pays for itself at discount rates of 1 to 7 percent, with the threshold moving from age 29 to 38 across that range. The same holds at marginal tax rates from 10 to 30 percent, and at every value of cumulative program spending we regard as defensible, including a strict upper bound that assigns every cohort a full four years of exposure. The MVPF becomes finite only at a discount rate of 10 percent or above, or at cumulative spending more than six times our central estimate.

Table C3: Marginal Value of Public Funds by earnings projection horizon

Earnings projected through age	PV of earnings gain per \$1 (1)	Additional tax revenue per \$1 (2)	Net cost to government per \$1 (3)	Beneficiary WTP per \$1 (4)	MVPF (5)
25 (no projection)	\$0.82	\$0.16	\$0.84	\$0.67	0.79
26	\$1.68	\$0.32	\$0.68	\$1.36	1.99
27	\$2.60	\$0.49	\$0.51	\$2.10	4.13
28	\$3.54	\$0.67	\$0.33	\$2.87	8.70
29	\$4.53	\$0.86	\$0.14	\$3.67	25.51
30	**\$5.53**	**\$1.05**	**–\$0.05**	**\$4.49**	∞
31	\$6.55	\$1.24	–\$0.24	\$5.31	∞
35	\$10.76	\$2.03	–\$1.03	\$8.73	∞
45	\$21.40	\$4.04	–\$3.04	\$17.36	∞
55	\$29.97	\$5.67	–\$4.67	\$24.31	∞
65 (baseline)	\$35.01	\$6.62	–\$5.62	\$28.39	∞

Notes: Every column is expressed per \$1 of Reading First spending, following the normalisation in Hendren and Sprung-Keyser (2020). Each row stops the lifetime earnings projection at the stated age and assumes the earnings effect is zero thereafter; the first row credits only the single year of earnings we observe. Column (1) is the present value at age 7, the mean age of K–3 exposure, of the earnings gain accumulated through the stated age, discounted at 3%. Column (2) applies a marginal tax-and-transfer rate of 18.9%. Column (3) is one dollar of program cost less column (2). Column (4) is column (1) less column (2). Column (5) is column (4) divided by column (3), and is infinite once the net cost to the government turns negative, that is, once the program has paid for itself. The projection holds the treatment effect at a constant percentage of the age-earnings profile of the 2015 American Community Survey, anchored to our own age-25 estimate of \$554 on a counterfactual mean of \$32,500; it assumes 0.5% annual real wage growth and cumulative Reading First spending of \$397 per exposed student, 3 years of K–3 exposure at \$132.27 per student-year, all in 2025 dollars. The appendix details the construction and reports sensitivity to each assumption.