Decoding Success: Evidence from a Federal Phonics Initiative*

Alex Hollingsworth, Alberto Ortega, and Nikolaos Zirogiannis†

March 26th, 2026

Abstract

The U.S. spends more per pupil than nearly any other country, yet reading proficiency has been stagnant for three decades. 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 SD, with effects that persist through eighth grade and have positive spillovers to math test scores. Effects are largest for high-poverty schools. District expenditure data show that Reading First dollars were spent on reading materials, instructional supplies, and not just on expanding the instructional workforce/salary. An analysis of digitized exam questions indicates that Reading First students outperformed peers on longer, more complex items and questions requiring advanced decoding skills, a pattern consistent with improved foundational reading ability. Comparing Reading First to unrestricted revenue from local tax elections, we find that discretionary funds flow to salaries rather than instructional materials, suggesting Reading First's effects reflect its curricular mandate rather than additional resources alone.

JEL Classifications: A20, H75, I22, I24
Keywords: Literacy, Phonics, Reading First

*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.

†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

The United States spends more per pupil than nearly any other country, yet fewer than one in three fourth graders read proficiently (NCES, 2022). This disconnect between resources and outcomes is not an idle concern, as early reading gaps are predictive of later-life outcomes such as educational attainment, earnings, and economic mobility (Chetty et al., 2011). Reading gaps disproportionately affect disadvantaged students, as roughly 80 percent of low-income children are not proficient readers (NCES, 2024). In addition, a central puzzle is that many educational interventions produce short-term gains that fade with time, while fewer improve long-run outcomes. This stagnation of academic achievement amid rising expenditures highlights a critical need to understand not just how much we spend, but also how we allocate resources to generate both short- and long-run returns (Chetty et al., 2011; Jacob et al., 2010).

A large body of literature explores the education production function, which relates student achievement to a variety of educational and environmental inputs (Hanushek, 1979). Within the classroom, research highlights the critical roles of teacher quality, class size, and school funding, among other things (Hoxby, 1996; Krueger, 1999; Chetty et al., 2014b; Jackson et al., 2016). [1] Despite this large literature, the specific pedagogical methods used in the classroom are notoriously difficult to observe in administrative data and thus have received less attention from economists. 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).

Decades of research in cognitive psychology, ranging from laboratory experiments to brain imaging, have established that children learn to read most effectively through systematic phonics instruction, which teaches the mapping between letters and sounds (National Reading Panel, 2000; Rayner, 1998; Castles et al., 2018). Despite this consensus, many American classrooms have historically relied on "whole language" approaches that emphasize context clues and text immersion over explicit decoding. This pedagogical divide grew so entrenched it became known as the "Reading Wars" (Mathews, 1966; Pearson, 1989; Kogan, 2025). Recently, spurred by the renewed "science of reading" movement, over 40 states have enacted legislation to promote phonics-based instruction (Schwartz, 2022; Kogan, 2025). However, understanding how children learn to read is not the same as understanding how to design policy that changes classroom practice at scale. Whether phonics-based instruction can meaningfully and durably improve student outcomes when delivered through large-scale government mandates, rather than controlled research settings, remains an open question.

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 (RF) program. Established under the No Child Left Behind Act, Reading First distributed nearly $1 billion annually na-

[1] Beyond the classroom, numerous studies examine external factors ranging from socioeconomic status and pollution exposure to nutrition (Reber, 2010; Dahl and Lochner, 2012; Ang, 2021; Persico and Venator, 2021; Cabral et al., 2021; Hollingsworth et al., 2022).

tionwide to promote phonics-based instruction in K–3 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 poverty indicators, while 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 dispersed, grow during the funding period, and persist at a positive and statistically significant level even after the last Reading First dollar is spent. Moreover, the gains do not fade out at the individual level: students exposed to the program in early grades continue to outperform their peers through eighth grade. We also find positive effects on third-grade math scores – driven by math 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 results show similar results indicating that districts receiving more revenue per student experienced larger gains.

Consistent with the program's design, which prioritized high-poverty areas, effects are larger in schools with above-median shares of Title I students. The effects of Reading First are largest for schools with a larger share of Hispanic students, though we find positive effects for White and Black students as well. Reading First was also remarkably cost-effective. The cost of the program was roughly $100 per treated student-year, yielding over a 0.1 standard deviation improvement in reading test scores. This is a fraction of the $3,000 per pupil per year that Jackson and Mackevicius (2024) estimate is typically required for a comparable gain.

We explore several channels through which Reading First may have improved achievement. First, using detailed district expenditure data, we show that Reading First funds were spent primarily on reading materials, general instructional supplies, and staff wages, but were not spent on observably expanding the teaching workforce. Second, we find no evidence that Reading First districts attracted more or higher-quality teachers, as measured by experience,

education, or campus-level value-added. Third, we digitize and classify the content of third-grade standardized exams and show that Reading First students outperform their non-treated peers on questions with more words, more complex questions, items testing vocabulary, and broad comprehension questions. These findings are consistent with the hypothesis that phonics instruction improves decoding skills, which in turn help students engage with more demanding text. We also compare Reading First expenditures to a more general revenue shock, the passage of school district tax elections, and find that tax based increases in discretionary funds flow primarily to salaries rather than supplies like reading materials, suggesting that Reading First's achievement effects reflect its curricular mandate rather than a simple infusion of additional money.

Finally, in on-going work we are extending the analyses to see if Reading First's short-run achievement gains translate into lasting improvements in human capital. The administrative data allow us to follow the earliest Reading First-exposed cohorts through their late 20s to examine high school graduation, ACT/SAT scores, college attendance, the share of students enrolling in remedial coursework, college graduation rates, and earnings in the mid-twenties. While a complete analysis of long-run outcomes is ongoing, these data position our study to trace a specific literacy intervention from elementary school through early adulthood.

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 remarkably little attention in the economics literature. Prior work commissioned by the federal government found no measurable effect on student reading comprehension (Gamse et al., 2008b). 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 Kindergarten through third-grade 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. Coupled together this provides a more accurate and comprehensive assessment of Reading First true efficacy.

Second, we provide 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 improved National Assessment of Educational Progress (NAEP) scores, though Wainer et al. (2026) argues that concurrent retention policies complicate causal attribution. Novicoff and Dee (2023) 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.[2] 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, which complicates the interpretation of studies examining more recent reforms (Westall and Cummings, 2023; Novicoff and Dee, 2023).

Third, we contribute to the literature on the persistence and fade-out 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 *do not* fade out through eighth grade, 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 Reading First funds were spent buying instructional and reading materials rather than solely being used to expand the teaching workforce. 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. 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 — a design question that has received considerable attention in other policy domains, particularly health care and welfare (Baicker and Staiger, 2005; Knight, 2002), with more limited evidence in education (Gordon, 2004; Cascio et al., 2010). Central to that design question is whether funding tied to specific practices produces different effects than an equivalent

[2]Figure A1 shows Google Trends data for the search term "Science of Reading." Interest in the term became widespread only after 2017.

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), but less is known about whether the composition of spending matters and whether conditional grants tied to specific practices produce different effects than unconditional revenue increases. 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 stacked 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 expenditures, a finding with implications for the design of federal education transfers in a system where thousands of local districts retain substantial autonomy over how resources are spent (Hoxby, 2000; Gordon, 2004).

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 and math achievement. Section 6 explores mechanisms, including expenditure patterns, school and teacher characteristics, question decoding analysis, and compares Reading First expenditures to other exogenous sources of school funding. Section 7 concludes.

2 Background

2.1 The Reading Wars

Debates about how best to teach children to read date back to at least the early nineteenth century (Mathews, 1966). The central fault line has long separated phonics-based approaches, which emphasize explicit instruction in letter-sound correspondences, from methods that prioritize meaning and context.[3] By the late 1950s, phonics had gained the upper hand, bolstered by Rudolf Flesch's 1955 bestseller *Why Johnny Can't Read* and Cold War anxieties about American educational standards. That consensus eroded quickly, however, as "whole language" rose to prominence through the 1970s and 1980s. Whole language held that reading acquisition is as natural as learning to speak, and that immersion in authentic texts was preferable to systematic, teacher-led skill instruction. By 1989, a prominent literacy scholar, P. David Pearson described the spread of whole language as nothing short of an "epidemic" (Pearson, 1989).

The debate took on a partisan dimension, with phonics increasingly identified with the Republican Party and whole language with progressive Democrats. By the late 1990s, more than 65 percent of state legislators introducing phonics bills were Republicans (Paterson, 2000).

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

California provided the most visible cautionary tale. After the state moved aggressively toward literature-based instruction in the 1980s, its 1992 NAEP scores ranked near the bottom nationally, prompting a legislative overhaul that ultimately passed by near-unanimous margins and inspiring roughly three dozen states to pass similar legislation emphasizing phonemic awareness and explicit phonics instruction by 1997 (Kogan, 2025; Lemann, 1997).

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) released its findings in 2000 with a strong endorsement of systematic and explicit phonics instruction, along with emphasis on phonemic awareness, fluency, vocabulary, and comprehension. Whole-language advocates disputed the report's meta-analysis drew predominantly on randomized studies, excluding the qualitative research that dominated whole-language scholarship—but the panel's conclusions provided the scientific rationale for what would become Reading First.

2.2 National Reading First Initiative

The Reading First (RF) program was established in 2002 under the No Child Left Behind Act (NCLB) of 2001. Authorized at $900 million in its first year, RF 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 design of Reading First was heavily influenced by the emerging consensus on "scientifically based reading research", which highlighted phonics instruction (National Reading Panel, 2000). By statute, Reading First required schools to use curricula and teaching practices proven effective by rigorous research. Reading First was unprecedented in scale, and became the academic cornerstone of No Child Left Behind (U.S. Department of Education, 2002). At its peak, over 5,000 schools in over 1,800 districts participated, implementing comprehensive reading reforms with RF support (U.S. Department of Education, 2011).

The foundation of Reading First rested on five essential components of early literacy National Reading Panel (2000). These included developing phonemic awareness —recognizing and manipulating the individual sounds in spoken words —and phonics, which links those sounds to written letters to enable decoding and word recognition. The program also emphasized vocabulary development—building the words students need to listen, speak, read, and write effectively—and reading fluency, the ability to read quickly and accurately as a bridge to comprehension. Reading First also highlighted reading comprehension strategies that emphasize understanding, remembering, and communicating with others about what has been read. RF required that local programs provide explicit, systematic teaching in each of the five essential areas. The federal guidance stressed that all instructional materials, lesson designs, and assessments should be grounded in scientifically proven approaches to teaching these five components (U.S. Department of Education, 2002). The law's criteria reshaped many states' and districts' approaches to selecting reading curricula, moving away from unproven methods

toward approaches validated by experimental research.

Beyond curriculum and instruction, RF funds supported extensive teacher professional development and coaching. School districts hired full-time reading coaches and provided training workshops to help K–3 teachers implement scientifically based reading practices with fidelity. The program also emphasized diagnostic assessments: schools were expected to use valid, reliable screening and progress-monitoring tests to identify struggling readers early and to guide instructional adjustments (U.S. Department of Education, 2002). Students who fell behind in any of the five key components were to receive immediate, research-based interventions.

Reading First was administered as formula grants to states, after each state educational agency submitted a detailed plan. Each state's Reading First application and plan required establishing a peer-review panel (of reading experts) to vet local district applications. 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. Local school districts were expected to adopt reading curricula, materials, and assessments aligned with scientifically based reading research, and to use funds for activities like teacher training and hiring reading coaches.

NCLB 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 RF between 2004 and 2007 of 17 school districts across 13 states found that the program led to changes in teaching practices—schools receiving RF funds spent significantly more time on the five key components of reading and provided more teacher training and coaching—yet it did not produce measurable gains in students' reading comprehension by grade 3 (Gamse et al., 2008a).

A separate federal implementation study that expanded the evaluation to 24 states similarly reported that RF schools devoted more time to reading instruction and adopted the required instructional elements. For instance, this report found that nearly 100 percent of RF schools had a reading coach by the end of the program, compared to 57 percent of non-RF schools. However, as with the earlier report, there was limited evidence that RF schools improved reading achievement relative to non-RF 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. Further, the survey components of the report suggest spillovers across schools: districts that received RF funds may have set up workshops that teachers in all schools (independent of RF funding) may attend.

By 2008, the accumulating evidence prompted questions about whether an intensive focus on phonics and other basic skills was enough to boost reading comprehension. The combination of modest-to-null impact findings from studies like Gamse et al. (2008a) and the program's administrative troubles ultimately led to the end of Reading First.[4] In response to the poor

[4]Although accountability was stated to be included in RF's structure, the program's early administration was mired in controversy (McCallion, 2005). There were claims of insufficient transparency regarding states' require-

results and management scandals, Congress cut Reading First's funding by roughly 60% in 2008, and then eliminated it entirely in fiscal year 2009 (U.S. Department of Education, 2011; Klein, 2008; Robelen, 2011).

2.3 Reading First in Texas

Texas was awarded over $630 million that was spent over seven years to implement Reading First across the state. Consistent with the national priorities of the Reading First program, Texas centered its efforts on the science of reading and focused on five essential components of phonemic awareness, phonics, fluency, vocabulary development, and text comprehension. For districts that received Reading First funds, Texas developed materials and trained technical assistance providers to work with individual 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, 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 nearly exactly 80% of the federal money.

The grey solid line in Figure 1D indicates that by 2005, all school districts that were ever-awarded a sub-grant had received at least some money. Three years later, in 2008, all third classrooms in treated districts were "fully treated," in the sense that the third-graders then had been exposed to Reading First curricula interventions for their entire K-3 period. Figure 1E shows the geographic distribution of school districts that received RF funding, by Texas county.

ments for receiving RF funding, and some questions about conflicts of interest among officials and contractors who helped implement the program in its early years (Klein, 2008).

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). 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. 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). Attendance data are available for our entire sample period, while disciplinary records begin in 1999.

In ongoing work, we follow the earliest Reading First cohorts through their late 20s to examine whether short-run reading gains translate into lasting improvements in high school completion, college attendance, and early-career earnings. These results will 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. 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 will 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 6.2.

3.2 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). PEIMS collects audited financial data from all Texas school districts in a standardized chart-of-accounts format, with key dimensions including the fund code (source of funds), the function code (purpose of expenditure), and the object code (type of expenditure). These data are available for our full study period.

Crucially, 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. One caveat is that these classifications reflect how districts coded their purchases, not necessarily their content. A phonics workbook may be recorded as a reading material rather than a textbook, so we cannot determine from the accounting data alone whether a given purchase reflects a phonics-based approach or some other instructional philosophy. A final limitation is that expenditures are reported at the district level, so while we can observe how much a district spent and on what, we cannot observe how Reading First funds were allocated across individual campuses or employees within a district.

3.3 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. Because phonics instruction is fundamentally about decoding: mapping written symbols to sounds and building fluency with the structure of words, the composition of exam questions

themselves may matter for how Reading First-exposed students perform. A student who has received strong phonics instruction may be better equipped to parse a question that is longer or more complex, and may be better at questions that explicitly test vocabulary or require close engagement with individual words. Analyzing question-level variation, therefore, allows us to go beyond average treatment effects and ask whether Reading First impacts were concentrated in precisely the types of items where phonics skills are most likely to be activated. 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 A2 summarizes the years for which we were able to obtain valid released 3rd grade exams and the number of questions by year. Reading exams are available for 1999–2004, 2006, 2009, and 2013–2019; math exams are available for 1998–2004, 2006, 2009, 2013–2014, and 2016–2019. 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.[5] We digitize the full set of 15 reading and math exams and classify each question along four dimensions: the total number of words in the question-and-answer pair, whether the question asks students to define or identify the meaning of a specific word (a vocabulary item), whether the question requires students to read an extended passage and demonstrate broad comprehension and the share of words in a reading passage that required advanced phonics skills. Our TEA data record correct and incorrect responses at the student level, which we merge with the question classifications to estimate heterogeneous treatment effects across item types.

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) between 2000 and 2015. Given that the passage of an election generates an exogenous increase in district revenues unrelated to any targeted programmatic objective, 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 3.2) 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

[5]Specifically, released reading exams for 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.

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.[6] We summarize Reading First eligibility (RFE_d) for district d as follows

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

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

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

Where there are N districts (961 for the balanced panel and 1,127 in the unbalanced panel) and RF State Funds$_t$ is the amount of money the federal government provided to Texas for Reading First in year t. Eligible districts get the maximum of $170,000 or a share of Reading First funds in year t equivalent to their share of Title I allocation. Our main specification will report Grant$_{dt}$ per the number of K-3 students in the district in 2002, but we also consider a specification where we use the level value of grants. One limitation of our eligibility measure is that we cannot obtain data on all components of the eligibility formula for every year. Thus,

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

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 predicted Reading First dollars allotted to a district and a district's Reading First expenditures. There is a clear positive relationship between the two variables, irrespective of whether we consider total amounts (in million dollars) or dollars per student. Figure 1G shows a similar pattern for cumulative district receipts and expenditures. Figure A3 shows the trend in actual Reading First expenditures and our instrument for predicted receipts.

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, school 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 2A. 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.[7] Thus, we employ the following difference-in-differences (DiD) event-study framework:

$$y_{cg(d)t} = \alpha_{cg} + \gamma_t + \sum_{\substack{j=1995 \\ j\neq 2003}}^{2017} \beta_j RFE_d \times \mathbb{1}\{t-2004=j\} + \sum_{t\neq 2002} \psi_t \left(X_{cg}^{2002} \times \mathbb{1}\{\text{Year}=t\}\right) + v_{cg(d)t}, \tag{3}$$

[7]Figure A4 shows the raw trend in standardized test scores without setting 2002 as a relative year.

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. $\mathbb{1}\{t - 2004 = j\}$ is an indicator variable equal to one if the observation year is j years from the start of Reading First in 2004. We also include interactions between campus-grade demographics, X_{cg}^{2002} (percent black, percent Hispanic, percent Asian/native), in 2002, and year fixed effects. Our results are robust to excluding these interactions, the inclusion of time-varying demographic controls, or excluding all demographic controls. We do not include time-varying covariates in our main specification, as they may also be correlated with the treatment. For example, residential sorting of households in response to the Reading First program might affect the demographic composition of a district. Further justification for this choice is provided by recent evidence indicating that DiD regressions that include time-varying regressors may yield biased estimates (Caetano et al., 2022).

The coefficients of interest 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 $j < -1$. The event-study design supports the common trends assumption if these pre-period coefficients are statistically insignificant and close to zero. The coefficients, β_j for $j >= 0$, 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

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

where $Post_{2004}$ is an indicator equal to one for years after 2004. Given that Reading First funding ended in 2011, we also estimate a version of equation (4) that allows the treatment effect to differ across the two periods, replacing $\beta RFE_d \times Post_{2004}$ 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.

4.3 Two-Stage Least Squares

Given that we can observe Reading First 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. Thus, 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 between 2004 and year t. For ease of interpretation, when presenting the estimated effects on test scores, we scale θ by \$1000. All other variables are as specified in Section 4.2.

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 (RD) 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 also present specifications where estimation is restricted to districts near the 13.6 percent threshold, and find qualitatively similar results.

4.4 First-Stage and District Revenues

We begin by establishing that Reading First Eligibility does indeed increase Reading First funding. Table 1 presents the relationship between Reading First eligibility and Reading First-specific revenue. Columns (1) and (2) present our difference-in-differences estimates. Columns (3) and (4) split the post-RF 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. Columns (5) and (6) use our predicted Reading First dollars from equation (2) to estimate Reading First funding. The odd columns in Table 1 consider all school districts, and the even columns restrict 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) Reading First eligible districts received an additional

$48 per K-3 student post-2003 relative to non eligible districts. Column (2) suggests that the estimated effect is similar ($41 per K-3 student) when restricting the analysis to districts within five-percentage points of the 13.6 fail rate. Reassuringly, 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 dispersed, Reading First eligible districts received $98 per K-3 student. Columns (5) and (6) show the first stage result for our instrumental variables approach. In column (6), we find that an additional predicted Reading First dollar per K-3 student results in an additional $0.28 per K-3 student. These estimates are similar when considering the total number of students per district instead (see Table B1).

We also consider 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 would lend additional credibility to any effects that we find operating through the channel of improved reading instruction, the target of the Reading First program. We present the effects of Reading First on the non-Reading First revenue in Table B2. In 2003, the average non-Reading First district received about $\$34,000$ per K-3 student. In column (1), we see that Reading First eligible districts received an additional $164 per K-3 student (approximately 1.3% increase in total revenue), a finding that is not statistically significant. The effect is similar when restricting the analysis to school districts within five-percentage points of the 13.6 fail rate. In column (6), we see that an additional predicted Reading First dollar is associated with an additional $0.23 in district total revenue. To place this estimate into context an additional one-standard deviation increase in predicted Reading First revenue per K-3 student ($151) would result in an additional $108 of total non-Reading First revenue per K-3 student or a .3% increase.

5 Test Scores

5.1 Third-Grade

Figure 2B presents the pre- and post-treatment effects from equation (3), where the outcome is standardized third-grade reading test scores. Confidence intervals are constructed from robust standard errors, clustered at the school district level to address over-dispersion. 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 effects persist until 2013, one year after the last Reading First dollar is spent. After this, the effect decreases, but remains positive.

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. We find similar effects, although for the bandwidth model, the estimates do not become statistically significant until 2009.

Table 2 presents the average treatment effect on Reading First eligibility on third-grade reading scores. Column 1 reports the pooled difference-in-differences estimate, indicating that campuses in eligible districts improved by 0.11 standard deviations after the program began. Column 2 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 larger post-period effect is consistent with the event study estimates discussed above, which suggest that the effects of Reading First took several years to materialize fully. Columns 3 through 5 move to the intensive margin, estimating the effect of cumulative Reading First dollars per K-3 student on third-grade reading. Column 3 reports the OLS estimate using actual Reading First expenditures, which yields a coefficient of similar magnitude to the 2SLS estimate but is less precisely estimated. Column 4 reports the reduced form estimate using predicted cumulative RF dollars. Column 5 reports the 2SLS estimate, instrumenting actual cumulative RF expenditures with predicted RF dollars. The first-stage F-statistic exceeds the conventional weak-instrument threshold. As with our difference-in-differences model, we see that additional cumulative RF expenditures improve third-grade reading.

Next, we examine whether the Reading First effects on reading test scores persist beyond third grade for an individual student. Results are reported in Figure 3, where, contrary to Table 2, the analysis is conducted at the student-year (as opposed to the campus-year) level. Each of the six horizontal panels in Figure 3, show pooled difference-in-differences estimates for students that we can observe from third through eight grade. For example, the first (top) panel includes students that we only observe in third grade, independent of whether we can observe them in the data afterwards. The second panel consists of students we observe in grades three and four, and so on. In the bottom panel, our sample consists of students who are observed consistently in third through eighth grade. Since we use data up to 2017, the bottom panel captures students who attended third grade in years prior to (and including) 2012. The estimates of the top panel of Figure 3 are consistent with the campus-year aggregated estimates of column 1 in Table 2. All panels show that the effects of Reading First persist all the way through grade eight. The effect on grades six and eight reading test scores is less precisely estimated, but still positive. This is an important result that highlights the persistent effects of the treatment beyond third grade and addresses any concerns regarding the compositional balance of our main specification.

As an additional check, we also consider the effect that Reading First had on students who did not attend a Reading First district in grades K-3, but instead moved into a Reading First district in fourth grade. Results in Figure A5 show evidence of spillover effects of the treatment. The top panel of Figure A5 suggests that third graders in non-Reading First districts (who would later move to a Reading First district starting in grade four) performed worse than their peers in non-Reading First districts. Although we acknowledge that students who move may not be comparable to other students, this pattern is informative since it suggests the

absence of positive selection among movers. That is, if the point estimate of the first panel was positive it would indicate that third graders in non-Reading First districts that move to a Reading First district in grade 4, were already performing better than other non-Reading First third graders even though neither the treatment (non-Reading First third graders that move to a Reading First district in fourth grade) or control group (non-Reading First third graders that stay in non-Reading First districts) were exposed to Reading First in third grade. Note that this negative result for third graders persists throughout all six panels in Figure 3. Once those third graders from non-Reading First districts move into a Reading First district in grade four, they experience improved performance that is comparable to students who were directly exposed to Reading First as third graders. This result provides evidence of spillover effects. Even though those fourth graders did not experience the Reading First treatment in grades K-3, they still benefit simply by being in a classroom of students and teachers who received the treatment. Similar to the results illustrated in Figure 3, students who move into a Reading First district in grade four and stay in that district through middle school have higher reading scores than their non-Reading First peers in all subsequent grades. Finally, we examine the effect of Reading First on math in Figure A6. We find improvements in math test scores. 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 6.3 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 of standardized math test questions.

5.2 Internal Validity Checks

Figure 2D and Figure 4 present internal validity checks for both of our DiD and 2SLS models, respectively. Each row of each figure is a separate specification. The first row is our baseline estimate for each specification. Row 2 replaces our interacted demographic controls with time-varying demographics. Row 3 excludes all demographic controls. Row 4 adds additional controls for the share of students who are gifted, economically disadvantaged, ESL, gifted, and special education.[8] Row 5 removes the balanced campus restriction and includes all campuses, resulting in an unbalanced panel of schools. Row 6 expands the set of years included in the sample. Row 7 removes any student weights. Texas school districts lie within 20 Education Service Centers regions which provide services to improve school performance.[9] 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

[8]Texas labels students on free or reduced lunch as economically disadvantaged.

[9]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.

exams twice.[10] Row 10 excludes these exam transition years. Row 11 keeps years before 2012, when Reading First funding was available. For each of these specifications, we find similar positive effects on third-grade reading.

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 A8 plots our average treatment effect estimate from equation (4) for different bandwidths around the 13.6 fail rate cut-off. Similarly, Figure A7 plots our estimate form our 2SLS specification across different bandwidths. Overall we see that our results are positive irrespective of bandwidth. Figure A8 shows that our effects are stable for bandwidths of three or greater. For the 2SLS estimates in Figure A7 we find that the point estimates are similar independent of the bandwidth choice, although less precise for bandwidths below four.

5.3 Heterogeneity

Panel (a) of Figure A9 shows that our effects are consistent for Black, White, and Hispanic students. In panel (b), we see an upward trend for Asian and Native American test scores, but the results are noisy, likely due to small sample sizes. We also consider a specification where the outcome is the difference between Black and Hispanic test scores relative to those of white students. Given the similarity in the estimates across race, it is not surprising that Figure A10 shows no differences in the racial test score gap. In Figure A11 we see that positive effects of Reading First on reading are similar for both male and female students. Figure A12 plots our event-study estimates for grades 4-8. The effects take longer to materialize as the grade analyzed increases. We see the strongest effects for grade four. Results become significant for grades five through eight starting in the year 2012 and are strongest for grade five, followed by grades six and seven. We find the smallest effects in grade eight. The effects in higher grades take longer to emerge, consistent with those gains being driven by students who received Reading First exposure in grades K–3. For example, the first fully treated eighth graders were students who entered kindergarten in 2003 and received the Reading First treatment in all K-3 grades. Those students took the eighth grade reading exam in 2012, which is the year we see evidence of improved performance in that grade.

We also consider whether our main results differ by campus characteristics measured in 2002, the year used to construct the eligibility criteria. Figure A13 reports estimates from a series of split-sample specifications, where we divide campuses at the median of a given characteristic and estimate the model separately for campuses above and below that median. Beyond documenting heterogeneity in treatment effects, this exercise 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

[10] 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 (STAAR) in 2012.

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. The first four rows 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, 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 five and six 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. Together, these results are consistent with the program reaching the students it was designed to serve: Reading First had its largest effects in higher-poverty schools with larger shares of disadvantaged students, the populations most directly targeted by the funding formula. The next three rows consider Special Education, Gifted, and ESL enrollment shares. In each case, estimates for campuses above and below the median are similar in magnitude, with overlapping confidence intervals. Motivated by a growing literature documenting the effects of air pollution on both short-run and long-run student performance (Currie et al., 2009; Ebenstein et al., 2016; Hollingsworth et al., 2022; Persico and Venator, 2021; Jacqz, 2022), we also split campuses by median PM 2.5 concentration. The estimates are similar across the two groups, though we note that pollution exposure is correlated with economic disadvantage and demographic composition in this setting.

5.4 Disciplinary Outcomes

Panel (a) of Figure A14 plots the event study estimates of Reading First eligibility on disciplinary outcomes, which include in- and out-of-school suspensions and expulsions. Although we see a decrease, the point estimates are not statistically significant. Panel (b) of Figure A14 shows that the decrease is driven by in-school suspensions, although the effect is imprecisely estimated.[11] Figure A15 shows no effect of Reading First eligibility on campus-grade absences per student.

6 Possible Mechanisms

6.1 Expenditures

In Section 4.4 we find that the Reading First program increased total district Reading First specific revenue with limited effects on non-Reading First revenue. We now examine the effects of Reading First eligibility on Reading First specific expenditures. Given the positive effects

[11] We also see no effect on expulsions (results available upon request).

of Reading First on student test scores, it is of interest to know what the funds were allocated to. Table 3 examines different broad expenditure categories. Panel A presents the pooled DiD estimates. Eligible districts spent approximately $48 more per K-3 student from their Reading First fund relative to ineligible districts after the program began. The largest share of this spending went to salary, wages, and benefits ($27.84), followed by supplies and materials ($17.57). Districts also spent modestly more on professional purchased services ($2.90) and travel ($0.85). There is no significant effect on miscellaneous operating costs. Panel B separates the treatment effect into the period when Reading First funding was active (2004 through 2011) and the period after federal funding ended. During the program years, eligible districts spent roughly $97 more per K-3 student from the Reading First fund, again concentrated in payroll and supplies. After 2011, these expenditures are much smaller or near zero across all categories, consistent with the end of Reading First funding. Panel C reports 2SLS estimates using predicted Reading First dollars per student as the instrument. An additional dollar of predicted Reading First funding translates to approximately $0.99 in total tracked Reading First expenditures, suggesting that nearly all of the predicted funding flowed through to actual spending. Of each predicted dollar, about 49 cents went to payroll, 31 cents to supplies and materials, and the remainder to professional services, travel, and other costs.

Given that supplies and materials represent the second largest expenditure category, we decompose this spending further. Table 4 breaks supplies and materials into their individual components, namely, books, general supplies, and testing materials. The $17.57 per-student increase in supplies and materials is split roughly evenly between books ($8.80) and general supplies ($8.59). There is also a small and statistically significant effect on testing materials ($0.17). Panel B separates spending into the active and post-funding periods. During 2004 through 2011, eligible districts spent about $17.90 more on books and $17.13 more on general supplies per K-3 student, with no significant effects after the funding ended. The 2SLS estimates in Panel C indicate that of each predicted Reading First dollar allocated to supplies and materials, roughly 10 cents purchased books, 21 cents went to general supplies, and a negligible share covered testing materials.

Table 5 further disaggregates books into its individual components, namely, reading materials, library books, and textbooks. Reading materials account for nearly all of the book spending, with a difference-in-differences estimate of $8.29 out of the total $8.80 books effect. Library books and textbooks are not individually statistically significant. Panel B indicates that during the active funding years, eligible districts spent about $16.73 more per K-3 student on reading materials, compared to $0.26 on library books and $0.90 on textbooks, with the latter two only marginally significant. After 2011, spending on all book subcategories is not significantly different from zero. The 2SLS estimates in Panel C indicate that of each predicted dollar allocated to books, roughly 8 cents went to reading materials and about 1 cent each to library books and textbooks. The limited effect on library books and textbooks may not be surprising given that the program's focus was on teaching students how to read, which presumably includes existing books. Library books and textbooks would not become obsolete because

a district joined Reading First. Thus, these results are consistent with phonics-based materials responding through the reading materials category. Although the payroll channel is a plausible mechanism through which Reading First affected student achievement (Loeb and Page, 2000; Hendricks, 2014; Biasi, 2021), a non-trivial share of spending also went to books, of which nearly all concentrated in reading materials. The reading materials category is likely where phonics-related curricula would be cataloged. Given that Reading First required phonics-based instruction, the additional spending on reading materials likely contributed to the observed test score effects. Another possibility is that schools and districts used the funds to update reading materials. Thus, the reading improvements we find may be due to updated or new materials independent of the curricular content. However, Panel B of Tables 3, 4, and 5 shows that these expenditures were not sustained beyond 2011, when RF funding ended. Yet, we find that the third-grade test score effects persist after 2011, suggesting that the instructional strategies and pedagogy associated with Reading First also contributed to student achievement independent of direct funding.

6.2 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 allow us to link students to teachers through classroom identifiers only after 2011 (the year Reading First funds ended). As an alternative, we construct various measures of campus (school) value added (Foy, 2024; Bacher-Hicks and Koedel, 2023; Biasi, 2021; Chetty et al., 2014a). We 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. This limitation creates a challenge that is well known in the value-added literature, where lagged scores are the standard control for prior student achievement (Bacher-Hicks and Koedel, 2023). We address this by constructing three versions of the campus value-added measure, each using a different set of controls in a first-stage student-level regression. The first version regresses standardized third-grade reading scores on student demographics (indicators for race, gender, economic disadvantage, ESL status, gifted status, special education, and immigrant status) and the student's second-grade attendance rate. Missing demographic indicators are coded to a common value, which functions as a missing indicator in the regression. This specification is the most conservative in that it does not control for any measure of prior academic achievement. The second version adds the student's third-grade math score to

the same set of controls. The math score serves as a same-year, other-subject proxy for student ability, partially addressing the absence of a lagged reading score. Including math scores helps control for differences in student ability across campuses that are not captured by the demographic and attendance controls alone. The last version constructs a fourth grade value added measure, where a lagged own-subject score (in third grade) is available. Here, we regress standardized fourth-grade reading scores on fourth-grade student demographics, third-grade attendance, and the student's third-grade reading score. This specification most closely resembles the standard value-added model used in the literature and provides a useful benchmark against the two third-grade measures.

For each of the three value added measures, 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. Given that these measures can be noisy, particularly for smaller campuses where a handful of students can shift the average substantially, we apply an empirical Bayes shrinkage procedure to improve the precision of the estimates (Kane and Staiger, 2008).[12]

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 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. We then follow those teachers over time, regardless of whether they remain at the same campus or move elsewhere. In any post-2003 campus-year, we can compute the average pre-period value added of the teachers currently at that campus by averaging the scores they carry from their 2002 assignments. 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). We then estimate the effect of Reading First eligibility on these campus-year means using our DiD , separate periods (during and after Reading First funding), and 2SLS specifications. Figure 5 reports the results. Across all four groups of teachers, all three value-added models, and all estimation approaches, the point estimates are centered around zero, and none are statistically significant. Reading First eligibility did not change the average pre-period quality of teachers at a campus, whether we look at the full set of classroom employees or decompose by entry, exit, or retention. These findings suggest that the reading achievement effects we document are not driven by a reshuffling of higher- or lower-quality teachers across RF and non-RF 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 A16. Again, we do not find any convincing evidence of changes in the

[12]The procedure decomposes the total variance in campus-year means into a signal component reflecting persistent campus quality, a sampling noise component that is inversely related to campus size, and a drift component estimated from the covariance of consecutive-year means. Each campus's value added is then shrunk toward the grand mean in proportion to its noise share, so that campuses with fewer students or more volatile scores receive more shrinkage. The resulting estimates provide a more reliable measure of underlying campus quality for use in our heterogeneity analysis.

composition of teacher degree attainment in RF districts. There is some suggestive evidence of a decrease in the share of teachers with no degrees and an increase in the share with a bachelor's degree, particularly for employees entering the classroom.

We next examine whether Reading First affected teacher staffing levels. Figure 6 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. As with the value-added analysis, we decompose the K-3 measures by entering, exiting, and incumbent classroom employees. We also examine nonclassroom employee FTE per 100 students. The DiD estimates are small and not statistically significant across all staffing measures, suggesting that Reading First eligibility did not generate large changes in the number of teachers per student on average. The 2SLS estimates, which scale the reduced form by the first stage, point to modest positive effects on K-3 FTE and K-3 Reading/ELA FTE per student, though the confidence intervals are wide. There is some suggestion in the 2SLS that Reading First funding may have supported additional K-3 and reading-specific staff, which would be consistent with the expenditure results showing that a substantial share of Reading First dollars went to salary and wages. However, given the imprecision of these estimates, we interpret them cautiously. There is no evidence of effects on nonclassroom staffing, teacher turnover, or the overall teacher-to-student ratio.

Lastly, Figure A17 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. As previously shown in Table 3, we find that district-level Reading First-specific expenditures for salary and wages increased. However, since Reading First funds were distributed across a subset of campuses within a district and likely reached only a subset of staff within these campuses, this analysis is naturally attenuated relative to what we might find if we could link Reading First spending to individual employees. We report results separately for classroom employees overall, entering teachers, exiting teachers, incumbents, and nonclassroom employees. Across all five groups and all four estimation approaches, the point estimates are close to zero, and none are statistically significant. There is little evidence that Reading First eligibility led to overall higher or lower salaries for any category of staff.

6.3 Decoding Test Questions

As described in Section 3.3, 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 campus-year level as the outcome. Table 6 presents results from our main DiD specification. Column 1 pools math and reading exams together, column 2 reports results for math only, and columns 3-5 report results for reading exams.

Panel A of Table 6 replicates 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 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 math (reading) question, Reading First-eligible students scored 0.02 (0.01) percentage points higher than non-Reading First-eligible students.

Panel C examines reading comprehension in more detail. We first focus on questions that test students' vocabulary skills by asking them what specific words of a reading passage mean. Our results show that, while all students do better in vocabulary related questions Reading First-eligible students score 1.8 percentage points higher than non-Reading First-eligible students. Column 4 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 three percentage points lower on those broad comprehension questions when compared to other question types. However, Reading First-eligible students score 0.6 percentage points higher on these items *relative* to non-Reading First students.

Next we classify questions based on the number of phonetically regular words they contain. Phonetic regularity is measured along the Center for the Improvement of Early Reading Achievement (CIERA) decodability scale that ranges from 1 to 8, where 1 indicates a word that can be easily decoded using basic phonics skills (eg. me, we, be) while 8 pertains to words that are very hard to decode and contain irregular phonetic patterns (Hoffman et al., 2002). We collapse the CIERA decodability scale into three categories, namely "basic phonics" to capture words with a CIERA score between 1-4, "advanced phonics" for words with a CIERA score between 5-7, and "irregular" for words with a CIERA score of 8. Column 5 of Table 6 interacts the treatment with the share of advanced phonics skills words in each question and again finds a positive and significant interaction for Reading First-eligible students. Overall, our analysis of question types reveals that Reading First-eligible students do better in more complex and challenging questions. What is more, their performance carries over to math exams suggesting that more competent readers are better positioned to understand and respond to math questions that involve longer verbal guidelines.

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. Namely, 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. However, we find no such effects on test scores from tax elections. If instead the Reading First estimates are larger or differently distributed across spending categories, that would point toward the curricular and instructional changes accompanying Reading First as the operative mechanism. 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 stacked difference-in-differences model (Wing et al., 2024) using data on Texas school district tax elections between 2000 and 2015 merged to Texas PEIMS financial data (discussed earlier). Treated districts are those where a tax increase passed, and the comparison is to districts where a similar measure was on the ballot but failed. We use the year of the first successful tax increase as the treatment year. Figure A18 presents the event study for total per-pupil revenue following a successful tax election. Pre-trends are flat and close to zero, and revenue rises by roughly \$200–\$300 per pupil in the post-election period, providing a meaningful and well-identified revenue shock for comparison. Table B3 examines how these discretionary dollars are spent. We begin by establishing that school revenue increases by \$312 per student for the average district post tax election in panel A. That increase comes entirely from local revenue sources which is what we expect. In sharp contrast to Reading First, discretionary revenue increases are almost entirely allocated to salaries and wages, with no significant effect on supplies and materials. Specifically, column (1) indicates that the average tax election increases total school spending by \$183 per student of which \$158 go towards salaries and benefits. Our 2SLS estimates presented in column (2) suggest that every additional tax election dollar per student increases total school expenditure by \$0.59 and salaries and wages by \$0.51. Decomposing the expenditure effect further in Panels (C) and (D) of Table B3, 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 suggest 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 Conclusion

Reading proficiency is an important predictor of later in life achievement. Government spending has, over the years, invested substantial resources towards improving reading curricula in elementary schools, often with limited results. We study a policy intervention that successfully improved reading scores among third graders in Texas, sustained those improvements through later grades and had performance spillovers in math and potentially other subjects as well. 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 cost of $100 per student-year, Reading First achieved an increase of .1 SD 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. While our results on the short-run effect of Reading First and its mechanisms are convincing, our analysis has not yet considered whether those short-run benefits translate into future, later-in-life achievements. Our next steps include an examination of whether Reading First affects a series of post-secondary education (university enrollment and graduation rates) and labor market outcomes (employment status and earnings). This will better position our work to contribute to the literature examining long-term effects of early education interventions.

References

Abott, C., Kogan, V., Lavertu, S., and Peskowitz, Z. (2020). School district operational spending and student outcomes: Evidence from tax elections in seven states. *Journal of Public Economics*, 183:104142.

Alesina, A., Baqir, R., and Hoxby, C. (2004). Political jurisdictions in heterogeneous communities. *Journal of Political Economy*, 112(2):348–396.

Algozzine, B., Wang, C., and Violette, A. S. (2011). Reexamining the relationship between academic achievement and social behavior. *Journal of positive behavior interventions*, 13(1):3–16.

Álvarez Marinelli, H., Berlinski, S., and Busso, M. (2024). Remedial education: Evidence from a sequence of experiments in colombia. *Journal of Human Resources*, 59(1):141–174.

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

Bacher-Hicks, A. and Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. In *Handbook of the Economics of Education*, volume 6, pages 93–134.

Baicker, K. and Staiger, D. (2005). Fiscal shenanigans, targeted federal health care funds, and patient mortality. *Quarterly Journal of Economics*, 120(1):345–386.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., and Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, 21(2):55–97.

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

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

Boser, U., Chingos, M., and Straus, C. (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, M., Kim, B., Rossin-Slater, M., Schnell, M., and Schwandt, H. (2021). Trauma at school: The impacts of shootings on students' human capital and economic outcomes. Technical report, National Bureau of Economic Research.

Caetano, C., Callaway, B., Payne, S., and Rodrigues, H. S. (2022). Difference in differences with time-varying covariates. *arXiv preprint arXiv:2202.02903*.

Cascio, E. U., Gordon, N., and Reber, S. (2010). Paying for progress: Conditional grants and the desegregation of southern schools. *Quarterly Journal of Economics*, 125(1):445–482.

Cascio, E. U. and Staiger, D. O. (2012). Knowledge, tests, and fadeout in educational interventions. NBER Working Paper No. 18038.

Castles, A., Rastle, K., and Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological science in the public interest*, 19(1):5–51.

Cellini, S. R., Ferreira, F., and Rothstein, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *Quarterly Journal of Economics*, 125(1):215–261.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics*, 126(4):1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

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

Cunha, F. and Heckman, J. J. (2007). The technology of skill formation. *American Economic Review*, 97(2):31–47.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931.

Currie, J., Hanushek, E. A., Kahn, E. M., Neidell, M., and Rivkin, S. G. (2009). Does pollution increase school absences? *The Review of Economics and Statistics*, 91(4):682–694.

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

Ebenstein, A., Lavy, V., and Roth, S. (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.

Foy, M. (2024). Teachers' unions and student achievement: Evidence from wisconsin.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., and Unlu, F. (2008a). Reading first impact study: Final report. Technical report, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., and Unlu, F. (2008b). Reading first impact study. final report. ncee 2009-4038. *National Center for Education Evaluation and Regional Assistance*.

Gordon, N. (2004). Do federal grants boost school spending? evidence from title I. *Journal of Public Economics*, 88(9–10):1771–1792.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, 14(3):351–388.

Hendricks, M. 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, J. V., Sailors, M., and Patterson, E. U. (2002). Decodable texts for beginning reading instruction: The year 2000 basals. *Journal of Literacy Research*, 34(3):269–298.

Hollingsworth, A., Huang, J. M., Rudik, I., and Sanders, N. J. (2022). A thousand cuts: Cumulative lead exposure reduces academic achievement. *Journal of Human Resources*, pages 0222–12169R2.

Hoxby, C. M. (1996). How teachers' unions affect education production. *Quarterly Journal of Economics*, 111(3):671–718.

Hoxby, C. M. (2000). Does competition among public schools benefit students and taxpayers? *American Economic Review*, 90(5):1209–1238.

Hoxby, C. M. (2021). Advanced cognitive skill deserts in the united states: Their likely causes and implications. *Brookings Papers on Economic Activity*, Spring:317–351.

Jackson, C. K., Johnson, R. C., and Persico, C. (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. K. and Mackevicius, C. L. (2024). What impacts can we expect from school spending policy? evidence from evaluations in the united states. *American Economic Journal: Applied Economics*, 16(1):412–446.

Jacob, B. A., Lefgren, L., and Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, 45(4):915–943.

Jacqz, I. (2022). Toxic test scores: The impact of chemical releases on standardized test performance within U.S. schools. *Journal of Environmental Economics and Management*, page 102628.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Klein, A. (2008). House panel would kill reading first funding. *Education Week*. June 19.

Knight, B. (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.

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

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

Lemann, N. (1997). The reading wars. *Atlantic Monthly*, 280(5):128–134.

Loeb, S. and Page, M. E. (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, S., McNally, S., and Viarengo, M. (2018). Changing how literacy is taught: Evidence on synthetic phonics. *American Economic Journal: Economic Policy*, 10(2):217–241.

Martorell, P., Stange, K., and McFarlin, I. (2016). Investing in schools: Capital spending, facility conditions, and student achievement. *Journal of Public Economics*, 140:13–29.

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

McCallion, G. (2005). Reading first and early reading first: Background and funding. Technical Report RL31241, Congressional Research Service.

McIntosh, K., Horner, R. H., Chard, D. J., Dickey, C. R., and Braun, D. H. (2008). Reading skills and function of problem behavior in typical school settings. *Journal of Special Education*, 42(3):131–147.

Morgan, P. L., Farkas, G., Tufis, P. A., and Sperling, R. A. (2008). Are reading and behavior problems risk factors for each other? *Journal of learning disabilities*, 41(5):417–436.

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 (2022). Reading state achievement, grade 8 - the nation's report card. National Center for Education Statistics. Accessed: 2025-01-10.

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. National Center for Education Statistics.

Novicoff, S. and Dee, T. S. (2023). The achievement effects of scaling early literacy reforms. *Educational Evaluation and Policy Analysis*, page 01623737251349178.

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

Pearson, P. D. (1989). Reading the whole-language movement. *The Elementary School Journal*, 90(2):231–241.

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

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

Robelen, E. W. (2011). Budget cuts raise questions about federal commitment to literacy. *Education Week*. March 14.

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

Schwartz, S. (2022). Which states have passed 'science of reading' laws? what's in them? *Education Week*. Accessed: 2025-01-10.

Spencer, N. (2024). Comprehensive early literacy policy and the "mississippi miracle". *Economics of Education Review*, 103:102598.

Strunk, K., Wright, T., Kilbride, T., Zhu, Q., Cummings, A., West, J., Turner, M., and De Voto, C. (2021). Read by grade three: Year one report.

Texas Education Agency (2024). 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.

U.S. Department of Education (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, H., Grabovsky, I., and Robinson, D. H. (2026). On education miracles in general (and those in mississippi in particular). *Significance*, 23(1):32–36.

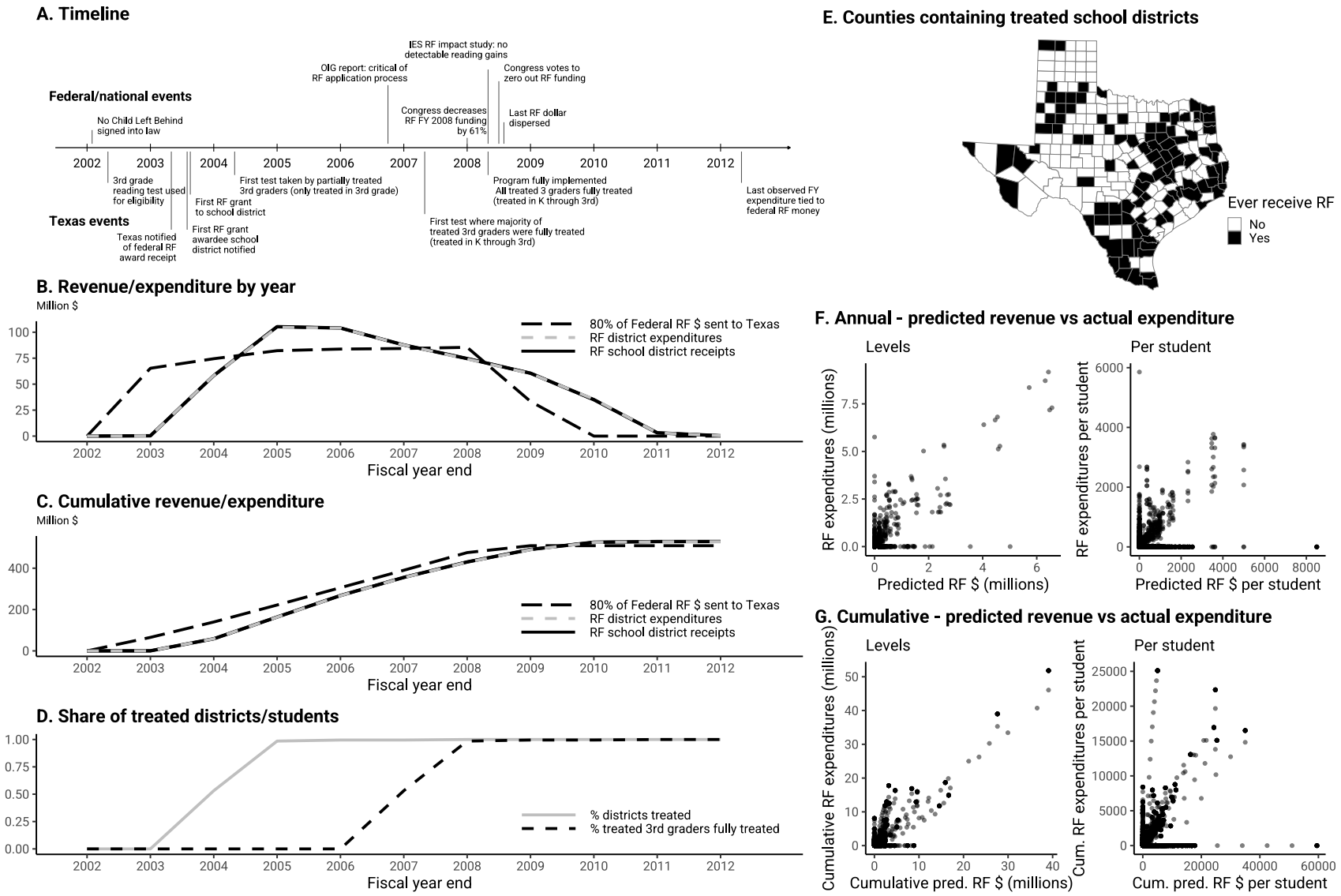
Westall, J. and Cummings, A. (2023). The effects of early literacy policies on student achievement. *Available at SSRN 4427675*.

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

Wing, C., Freedman, S. M., and Hollingsworth, A. (2024). Stacked difference-in-differences. Technical report, National Bureau of Economic Research.

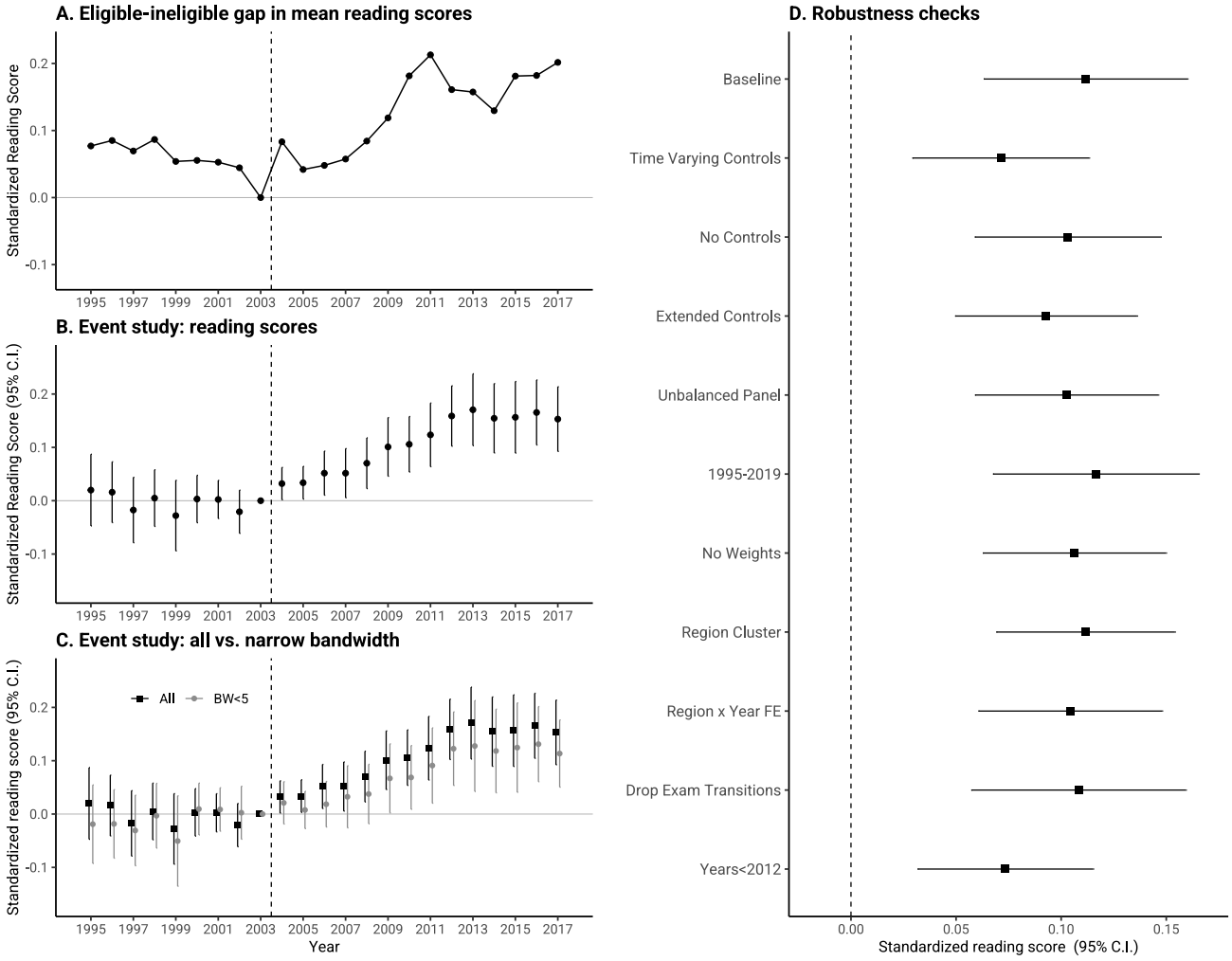
Xu, Z., Özek, U., Levin, J., and Hoon Lee, D. (2023). Effects of large-scale early math interventions on student outcomes: Evidence from kentucky's math achievement fund. *Educational Evaluation and Policy Analysis*, page 01623737241303142.

Figure 1: Overview of spatial and temporal sources of variation



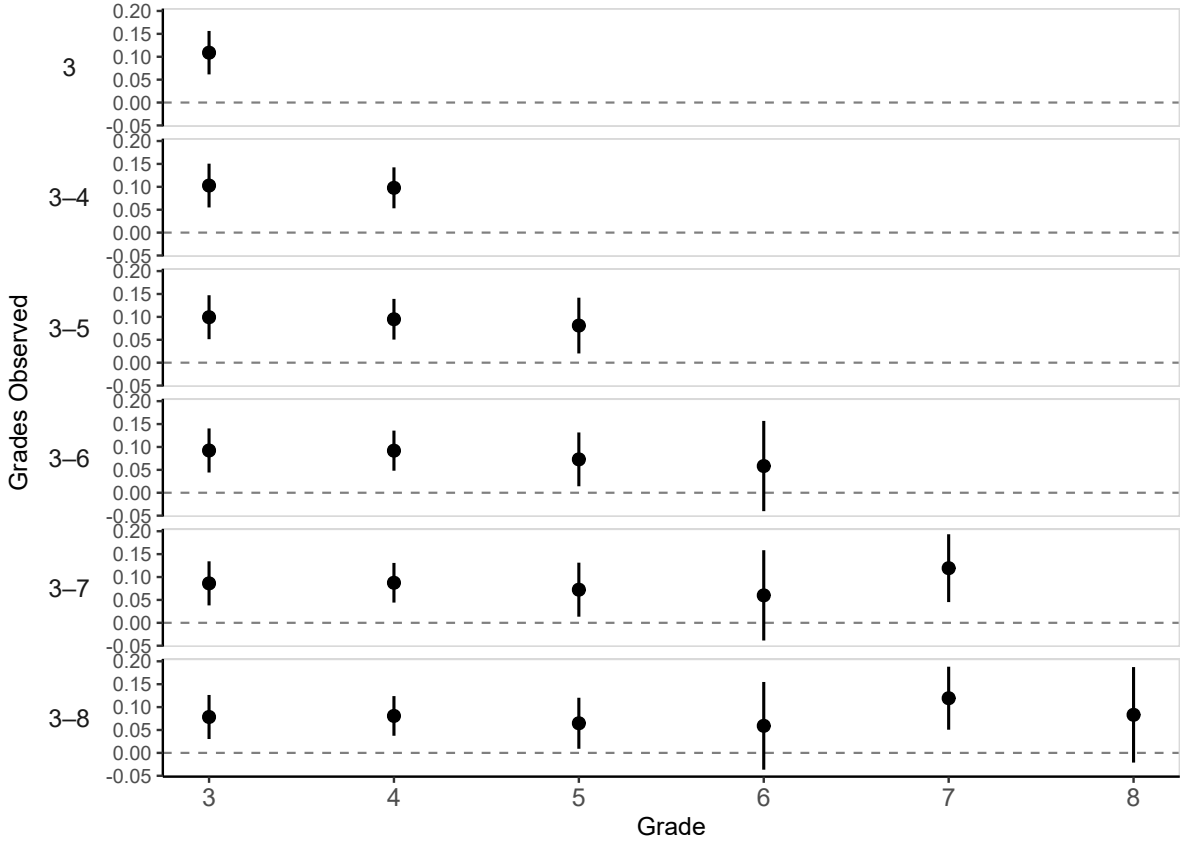
Notes: Panel A presents a timeline of the Reading First (RF) program in Texas alongside key federal events. Panel B plots total district RF receipts and expenditures by year. Panel C shows cumulative RF receipts and expenditures across years. Panel D displays the share of awarded districts that had received 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 and actual RF expenditures (first stage). Panel G shows cumulative predicted versus actual RF receipts.

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



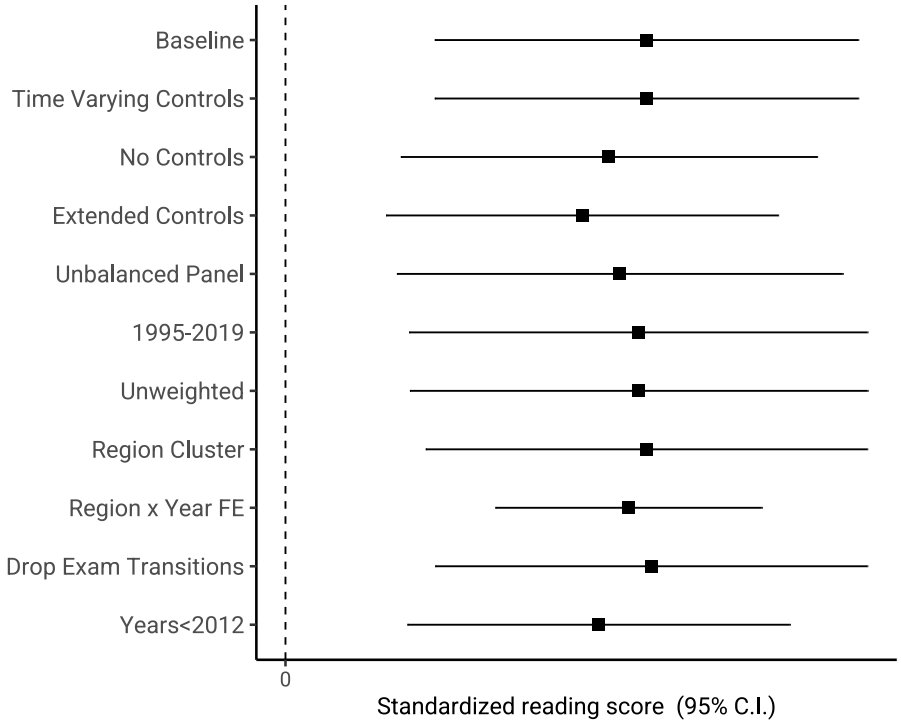
Notes: Panel A plots the raw difference in standardized third-grade reading scores between RF-eligible and non-eligible districts, relative to 2003. Panel B presents event-study estimates from equation (3) with 95% confidence intervals. Panel C restricts the sample to districts within five percentage points of the 13.6% fail rate eligibility threshold. Panel D reports specification checks (see Section 5.2 for details). All models include campus-by-grade and year fixed effects and demographic controls (share Black, Hispanic, Asian/Native in 2002) interacted with year indicators. Standard errors clustered at the district level.

Figure 3: Effect of Reading First eligibility on reading scores across multiple grades



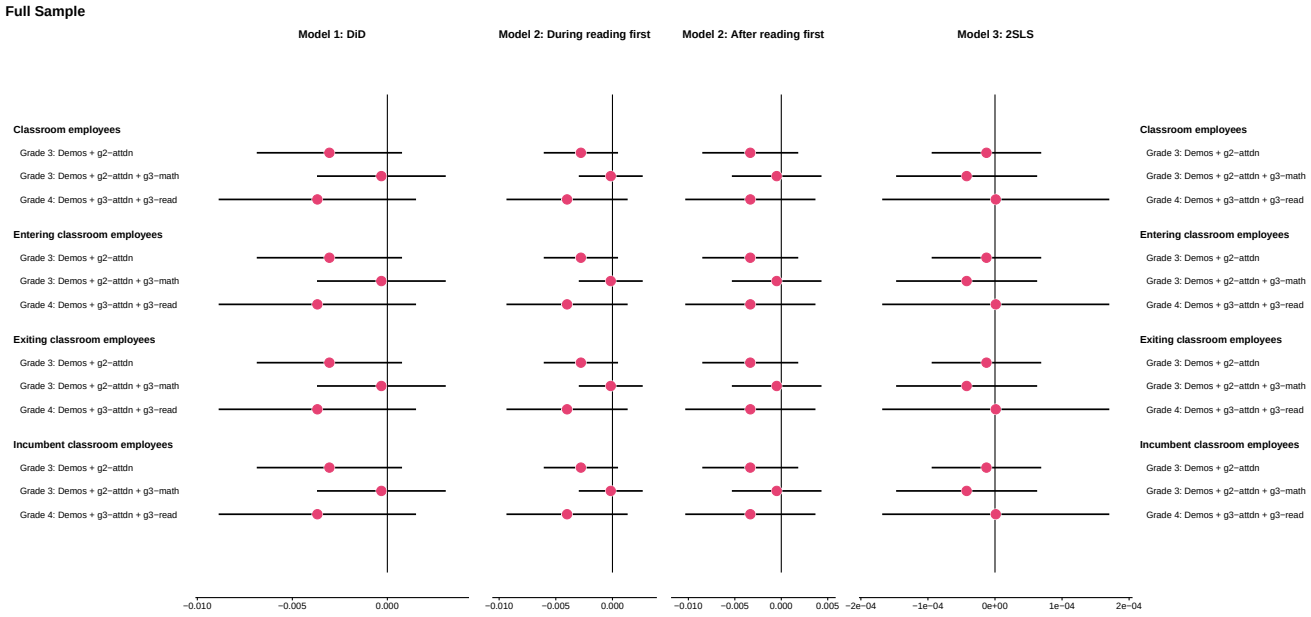
Notes: Each horizontal panel reports pooled difference-in-differences estimates at the student-year level for progressively longer windows of student observation. The top panel includes all students observed in third grade; the bottom panel restricts to students observed continuously from third through eighth grade. Bars represent 95% confidence intervals. All models include campus-by-grade and year fixed effects and demographic controls (share Black, Hispanic, Asian/Native in 2002) interacted with year indicators. Standard errors clustered at the district level.

Figure 4: Robustness of instrumental variables results



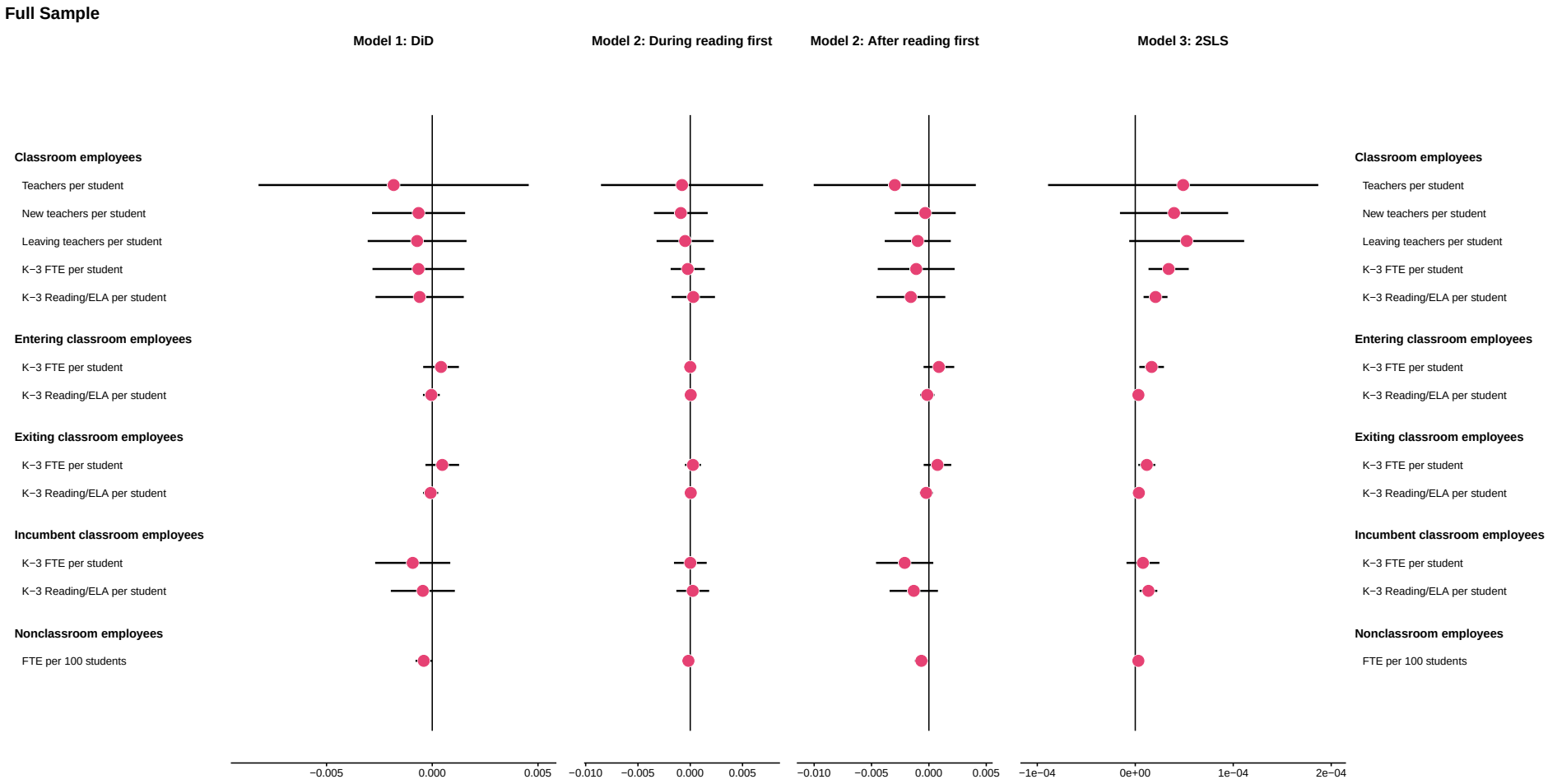
Notes: Each row reports the 2SLS estimate of cumulative RF funding per K-3 students on standardized third-grade reading scores under a different specification. The baseline specification includes campus-by-grade and year fixed effects and demographic controls (share Black, Hispanic, Asian/Native in 2002) interacted with year indicators. See text for details on each specification. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure 5: Changes in average teacher value added by campus



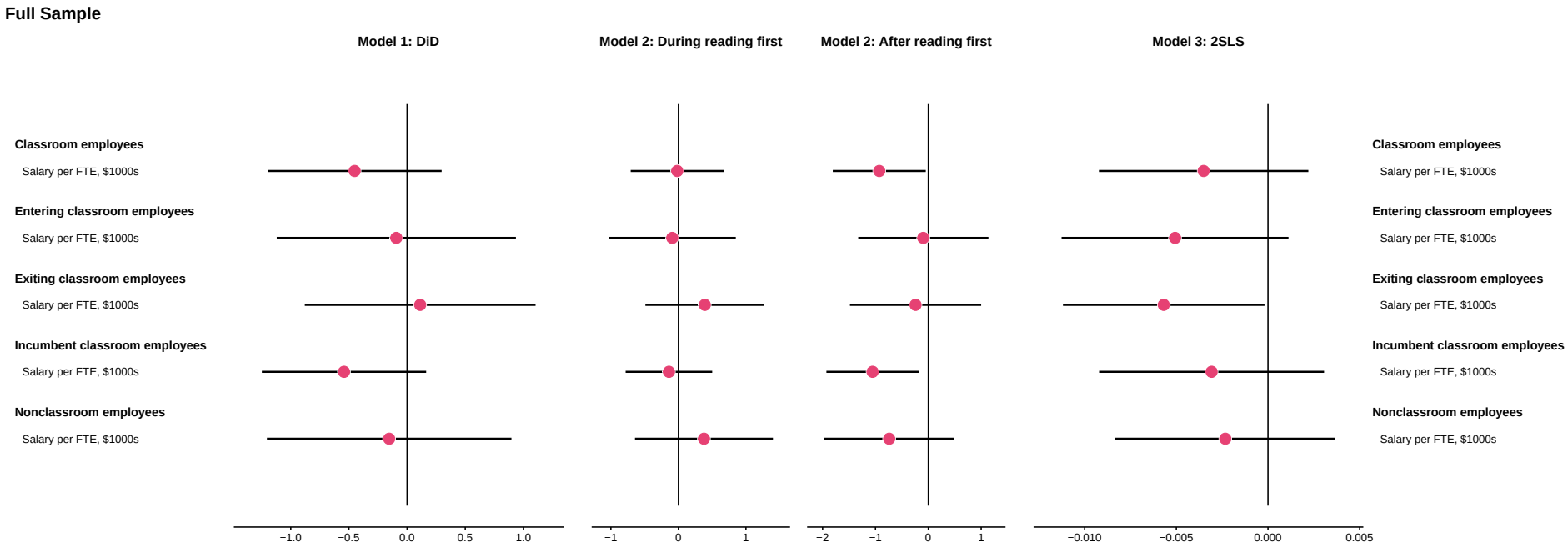
Notes: Each panel reports difference-in-differences, separate-period, and 2SLS estimates of the effect of RF eligibility on campus-level average pre-period teacher value added. Value added is estimated using student-level data from 1995–2002 under three specifications (see text). Teachers are assigned 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. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure 6: 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. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure 7: 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 teachers, exiting teachers, incumbents, and nonclassroom employees. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

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

	DiD		DiD, separate periods		First-stage	
	(1)	(2)	(3)	(4)	(5)	(6)
1(RF Eligible) x 1(year > 2003)	48.52***	41.01***				
	(7.74)	(9.50)				
1(RF Eligible) x 1(2004 ≤ year ≤ 2011)			97.70***	84.62***		
			(15.15)	(19.05)		
1(RF Eligible) x 1(year > 2011)			1.70	0.01		
			(1.36)	(1.24)		
Predicted RF per student					0.19***	0.28***
					(0.03)	(0.08)
					(0.03)	(0.08)
F-stat (instrument)					31.72	16.14
Observations	23,245	8,628	23,245	8,628	23,245	8,628
District FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Weights	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students
Bandwidth	None	5	None	5	None	5

Notes: Each column reports a separate regression of RF-specific revenue per K–3 student 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 periods. Columns (5)–(6): first-stage regression of actual RF revenue on predicted RF revenue per K–3 student. 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 K–3 enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

	DiD	DiD, separate periods	Endogenous	Reduced form	2SLS
	(1)	(2)	(3)	(4)	(5)
RF Post	0.112***				
	(0.025)				
RF Eligible × During RF		0.074***			
		(0.021)			
RF Eligible × After RF		0.162***			
		(0.033)			
Cumulative RF $/1 million (Endogenous)			0.005***		
			(0.002)		
Cumulative Predicted RF $/1 million				1.305***	
				(0.031)	
Cumulative RF $/1 million (2SLS)					0.005***
					(0.002)
First-stage F-stat					1,795
Observations	59,658	59,658	59,658	59,658	59,658
Campus FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Demographic Controls × Year	✓	✓	✓	✓	✓
Clustered SEs (District)	✓	✓	✓	✓	✓

Notes: The outcome variable is standardized third-grade reading scores. Column (1) reports the pooled difference-in-differences estimate. Column (2) separates the treatment effect into the during-RF (2004–2011) and post-RF periods. Column (3) reports OLS estimates using actual cumulative RF expenditures per K-3 students . Column (4) reports the reduced-form estimate using predicted cumulative RF funding (in millions). Column (5) instruments cumulative RF funding per K-3 students with cumulative predicted RF funding (in millions) . All models include campus and year fixed effects and demographic controls (share Black, Hispanic, Asian/Native in 2002) interacted with year indicators. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: 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
	(1)	(2)	(3)	(4)	(5)	(6)
A. DiD						
1(RF Eligible) × 1(year>2003)	48.04***	27.84***	0.85***	2.90***	17.57***	−1.12
	(7.63)	(3.70)	(0.25)	(0.73)	(3.11)	(2.32)
Observations	23,245	23,245	23,245	23,245	23,245	23,245
B. DiD, separate periods						
1(RF Eligible) × 1(2004≤year≤2011)	96.74***	55.72***	1.69***	5.74***	35.39***	−1.79
	(14.93)	(7.30)	(0.46)	(1.45)	(6.15)	(4.17)
1(RF Eligible) × 1(year>2011)	1.67	1.30*	0.06	0.20**	0.60	−0.49
	(1.36)	(0.72)	(0.07)	(0.09)	(0.45)	(0.57)
Observations	23,245	23,245	23,245	23,245	23,245	23,245
C. 2SLS						
Predicted RF $ per student	0.99***	0.49***	0.03***	0.06***	0.31***	0.10*
	(0.01)	(0.03)	(0.00)	(0.01)	(0.02)	(0.05)
F-stat (instrument)	41.93	41.93	41.93	41.93	41.93	41.93
Observations	23,011	23,011	23,011	23,011	23,011	23,011
District FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Weights	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students	K-3 students

Notes: Each column reports a separate regression of RF-specific expenditures per K–3 student by broad expenditure category. Panel A: pooled difference-in-differences. Panel B: separate treatment effects for the during-RF (2004–2011) and post-RF periods. Panel C: 2SLS estimates instrumenting actual RF expenditures per K–3 student with predicted RF funding per K–3 student. 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 4: Effect of Reading First eligibility on Reading First school district specific expenditure categories

	Supplies & materials	Books	General supplies	Testing materials
	(1)	(2)	(3)	(4)
A. DiD				
1(RF Eligible) × 1(year>2003)	17.57***	8.80***	8.59***	0.17***
	(3.11)	(2.15)	(2.11)	(0.06)
Observations	23,245	23,245	23,245	23,245
B. DiD, separate periods				
1(RF Eligible) × 1(2004≤year≤2011)	35.39***	17.90***	17.13***	0.34***
	(6.15)	(4.29)	(4.09)	(0.13)
1(RF Eligible) × 1(year>2011)	0.60	0.14	0.47	0.00
	(0.45)	(0.22)	(0.41)	(0.01)
Observations	23,245	23,245	23,245	23,245
C. 2SLS				
Predicted RF $ per student	0.31***	0.10***	0.21***	0.00***
	(0.02)	(0.02)	(0.02)	(0.00)
F-stat (instrument)	41.93	41.93	41.93	41.93
Observations	23,011	23,011	23,011	23,011
District FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Weights	K-3 students	K-3 students	K-3 students	K-3 students

Notes: Each column 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 periods. Panel C: 2SLS estimates instrumenting actual RF expenditures per K–3 student with predicted RF funding per K–3 student. 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 5: Effect of Reading First eligibility on Reading First school district book-related expenditure categories

	Books	Reading materials	Library books	Textbooks
	(1)	(2)	(3)	(4)
A. DiD				
1(RF Eligible) × 1(year>2003)	8.80***	8.29***	0.12	0.40
	(2.15)	(2.16)	(0.08)	(0.27)
Observations	23,245	23,245	23,245	23,245
B. DiD, separate periods				
1(RF Eligible) × 1(2004≤year≤2011)	17.90***	16.73***	0.26*	0.90*
	(4.29)	(4.29)	(0.16)	(0.55)
1(RF Eligible) × 1(year>2011)	0.14	0.24	−0.01	−0.09
	(0.22)	(0.20)	(0.01)	(0.06)
Observations	23,245	23,245	23,245	23,245
C. 2SLS				
Predicted RF $ per student	0.10***	0.08***	0.01**	0.01***
	(0.02)	(0.01)	(0.00)	(0.00)
F-stat (instrument)	41.93	41.93	41.93	41.93
Observations	23,011	23,011	23,011	23,011
District FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Weights	K-3 students	K-3 students	K-3 students	K-3 students

Notes: Each column reports a separate regression of RF-specific expenditures per K–3 student by book subcategory. Panel A: pooled difference-in-differences. Panel B: separate treatment effects for the during-RF (2004–2011) and post-RF periods. Panel C: 2SLS estimates instrumenting actual RF expenditures per K–3 student with predicted RF funding per K–3 student. All specifications include district and year fixed effects, weighted by K–3 enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

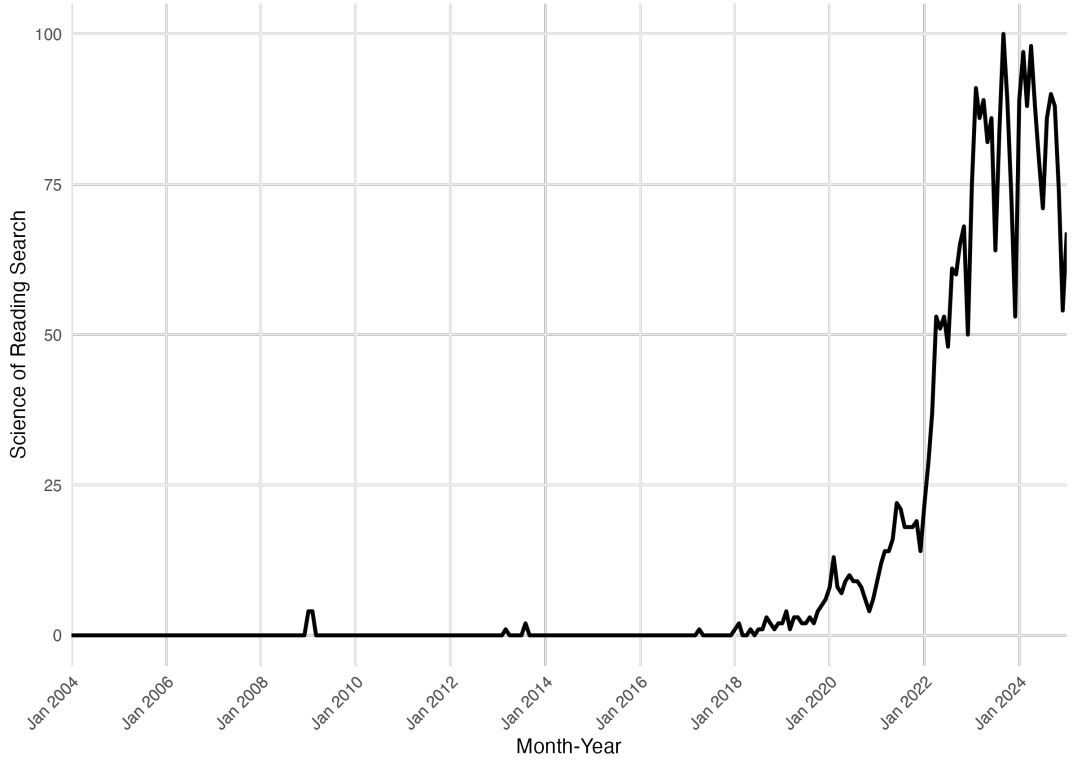
Table 6: Third-grade share of correct responses by Reading First eligibility

	Pooled	Math	Reading		
	(1)	(2)	(3)	(4)	(5)
Panel A. Baseline DiD					
1(RF Eligible) × 1(year>2003)	1.479***	1.613***	1.238***		
	(0.5228)	(0.6027)	(0.4157)		
Observations	8,622,308	4,412,914	4,209,394		
R^2	0.33873	0.27317	0.46368		
Panel B. DiD × Word Count					
1(RF Eligible) × 1(year>2003)	0.7594	0.7571	0.7113*		
	(0.4884)	(0.5739)	(0.4058)		
Word Count	−0.0888***	−0.0753***	−0.1021***		
	(0.0019)	(0.0026)	(0.0013)		
1(RF Eligible) × 1(year>2003) × Word Count	0.0179***	0.0214***	0.0130***		
	(0.0023)	(0.0027)	(0.0019)		
Observations	8,622,308	4,412,914	4,209,394		
R^2	0.35147	0.28109	0.48289		
Panel C. Reading Comprehension					
1(RF Eligible) × 1(year>2003)			0.9496**	1.108***	0.5870
			(0.3769)	(0.4120)	(0.4709)
Vocabulary			2.440***		
			(0.1006)		
1(RF Eligible) × 1(year>2003) × Vocabulary			1.847***		
			(0.1187)		
Broad Comprehension				−2.947***	
				(0.0464)	
1(RF Eligible) × 1(year>2003) × Broad Comprehension				0.5824***	
				(0.0747)	
Pct Advanced Phonics					0.0559***
					(0.0018)
1(RF Eligible) × 1(year>2003) × Pct Advanced Phonics					0.0443***
					(0.0055)
Observations			4,275,034	4,209,394	4,205,953
R^2			0.46458	0.46853	0.46372
Campus FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓

Notes: The outcome is the share of questions answered correctly at the campus-year-grade-question level. Column (1) pools math and reading exams; column (2) reports results for math only; columns (3)–(5) report results for reading exams. Panel A: baseline difference-in-differences. Panel B: interacts the treatment with the word count of each question. Panel C (reading exams only): column (3) interacts with an indicator for vocabulary questions, column (4) with an indicator for broad comprehension questions, and column (5) with the share of advanced phonics words. All specifications include campus and year fixed effects. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

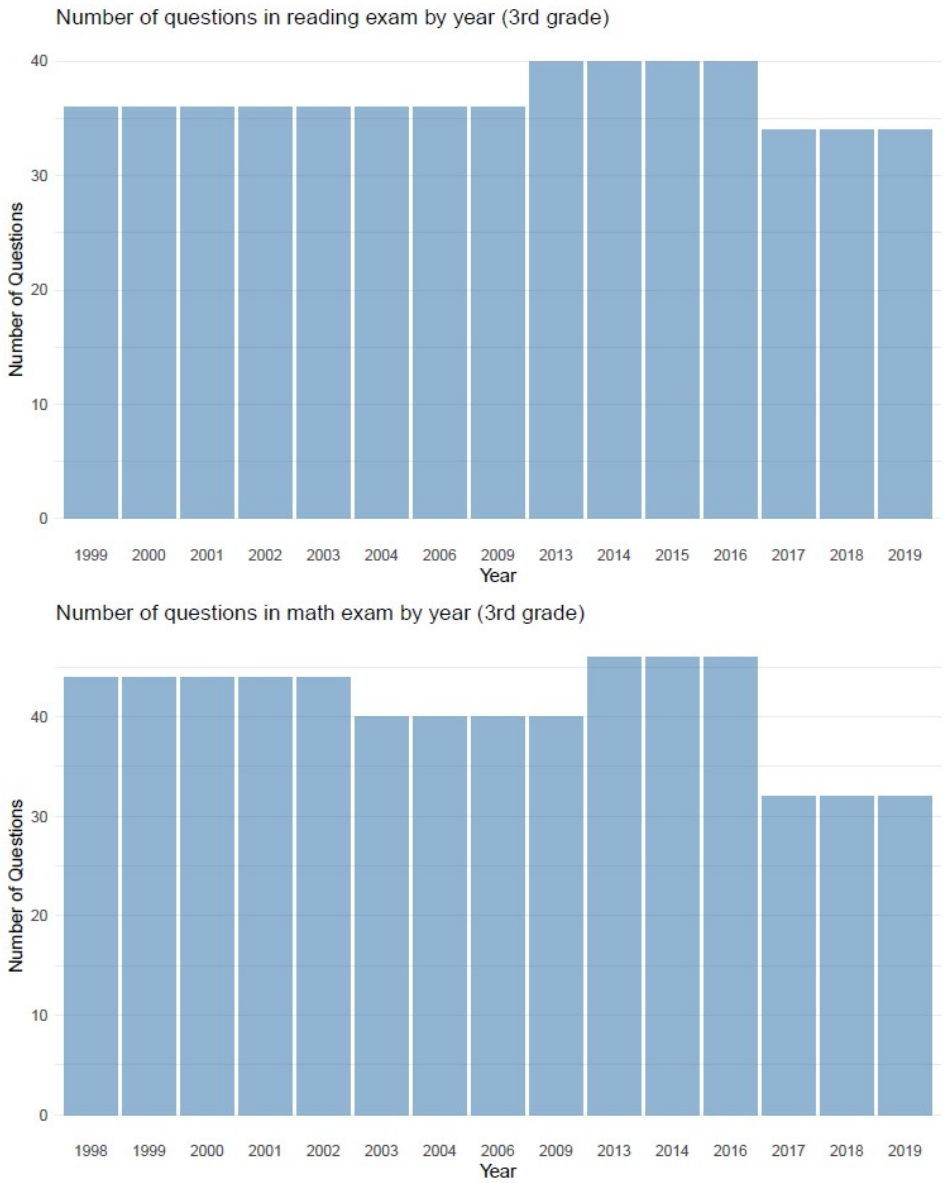
Appendix

Figure A1: Google search trend for "Science of Reading"



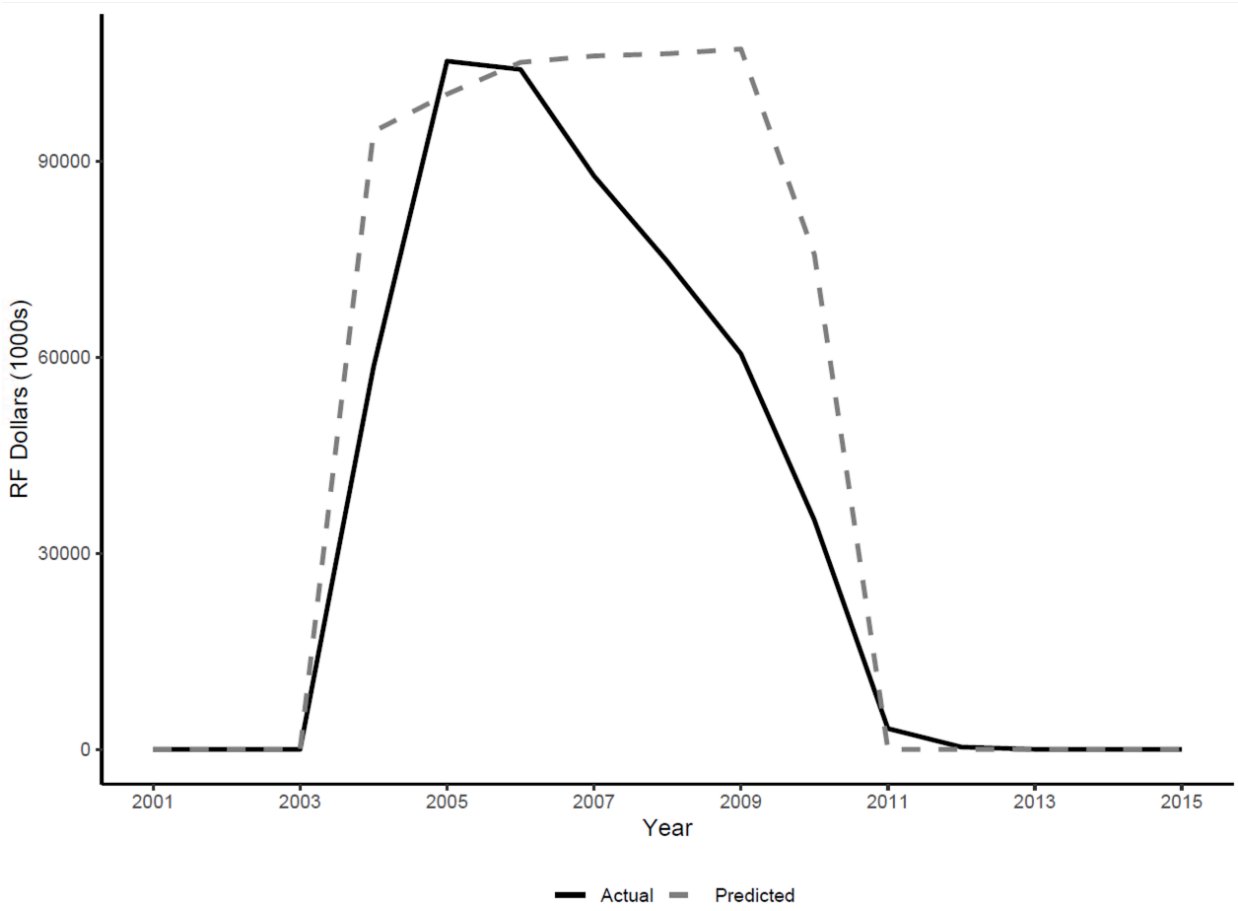
Notes: Google Trends index for the search term "Science of Reading," normalized to 100 at peak interest. The figure shows that widespread interest in the term emerged only after 2017, well after the Reading First program period.

Figure A2: Number of questions for third-grade reading and math exams by year



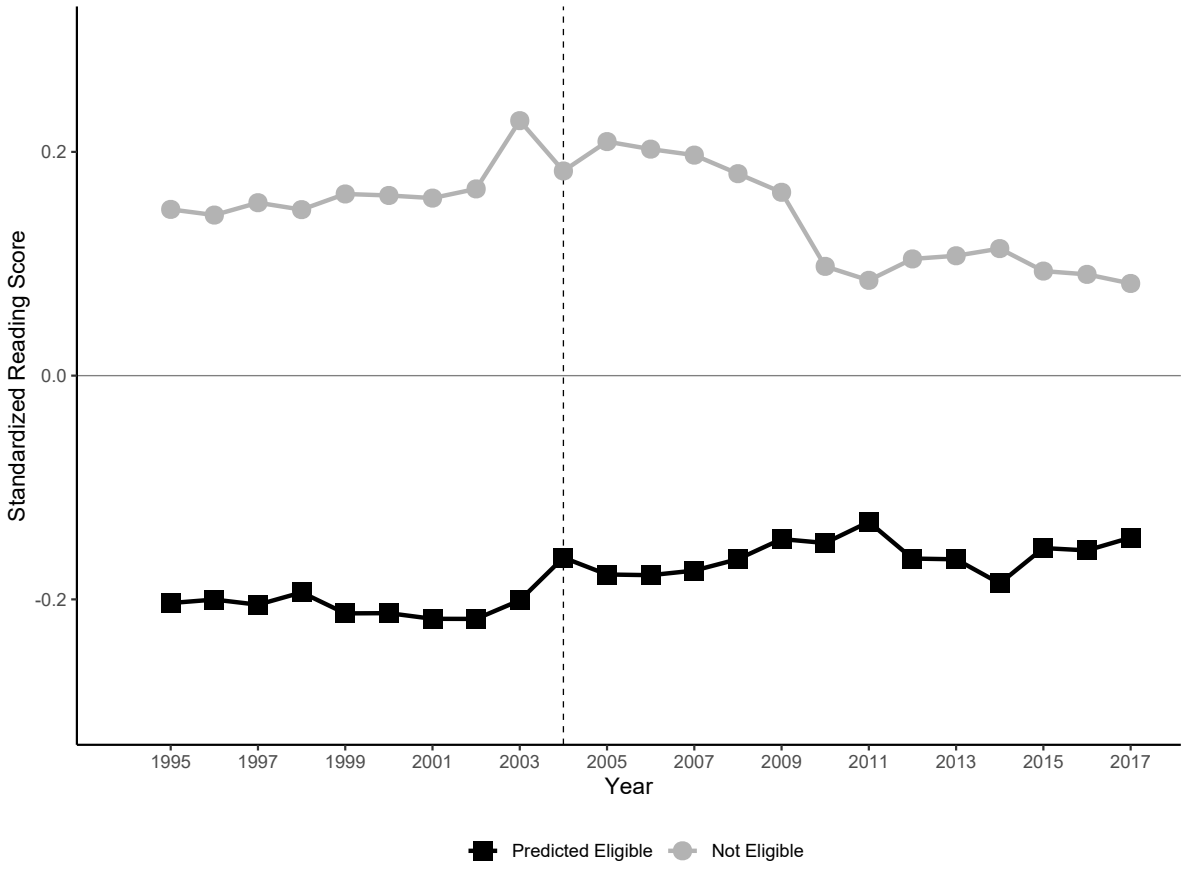
Notes: Number of released third-grade reading and math exam questions available by year. Exams are drawn from the TAAS (1995–2002), TAKS (2003–2011), and STAAR (2012–present) testing programs. See Section 3.3 for details on exam availability.

Figure A3: Reading First funds by year



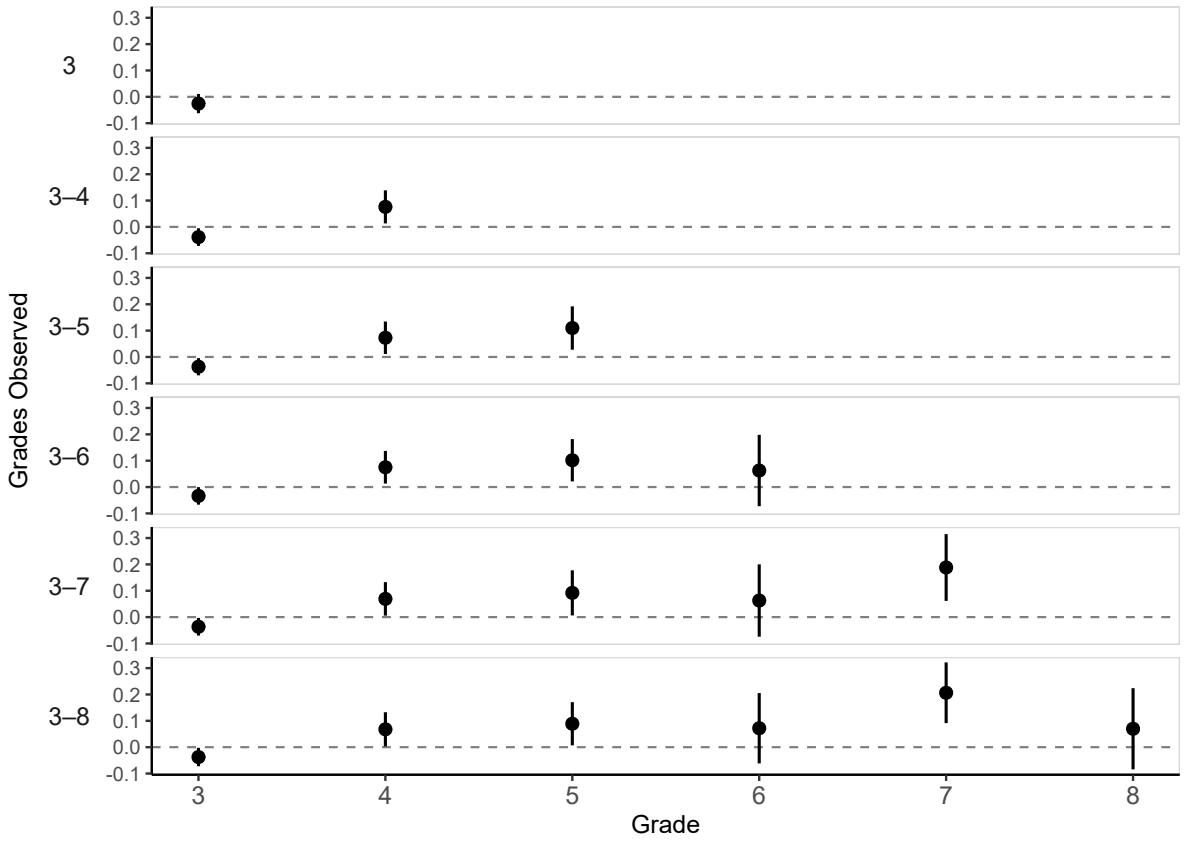
Notes: Actual RF expenditures and predicted RF receipts (from equation 2) by year for Texas school districts.

Figure A4: Trend in third-grade reading by Reading First eligibility



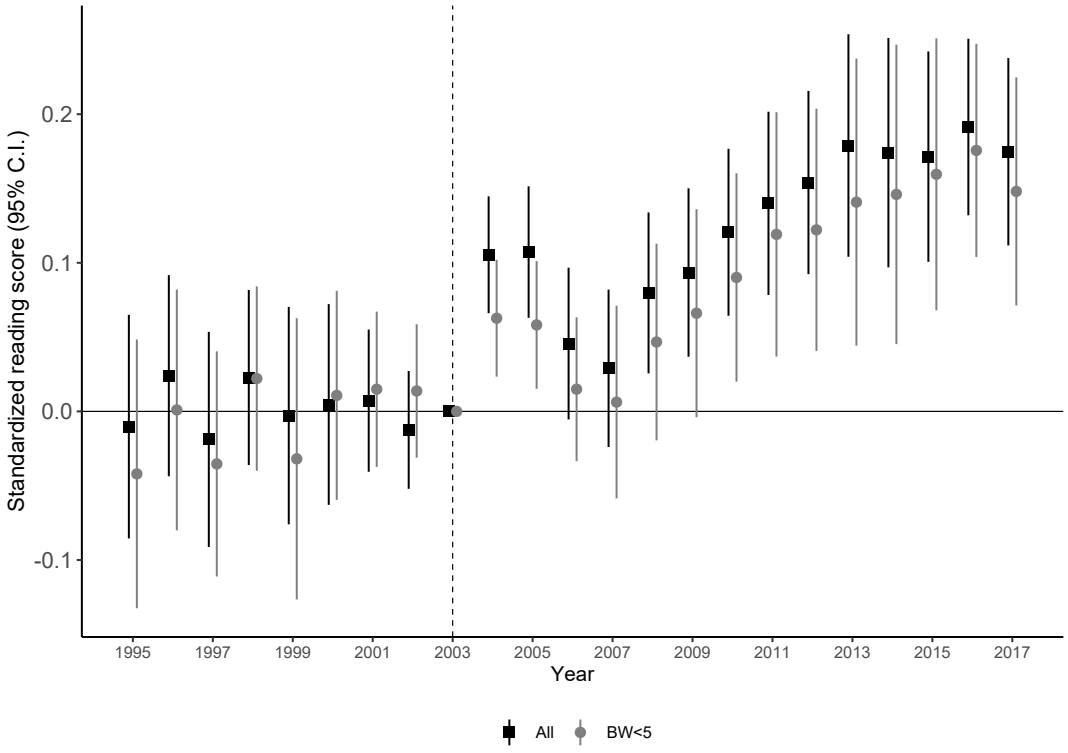
Notes: Raw standardized third-grade reading scores over time for RF-eligible and non-eligible districts.

Figure A5: Effect of Reading First eligibility on reading scores for students who move into a Reading First district in fourth grade



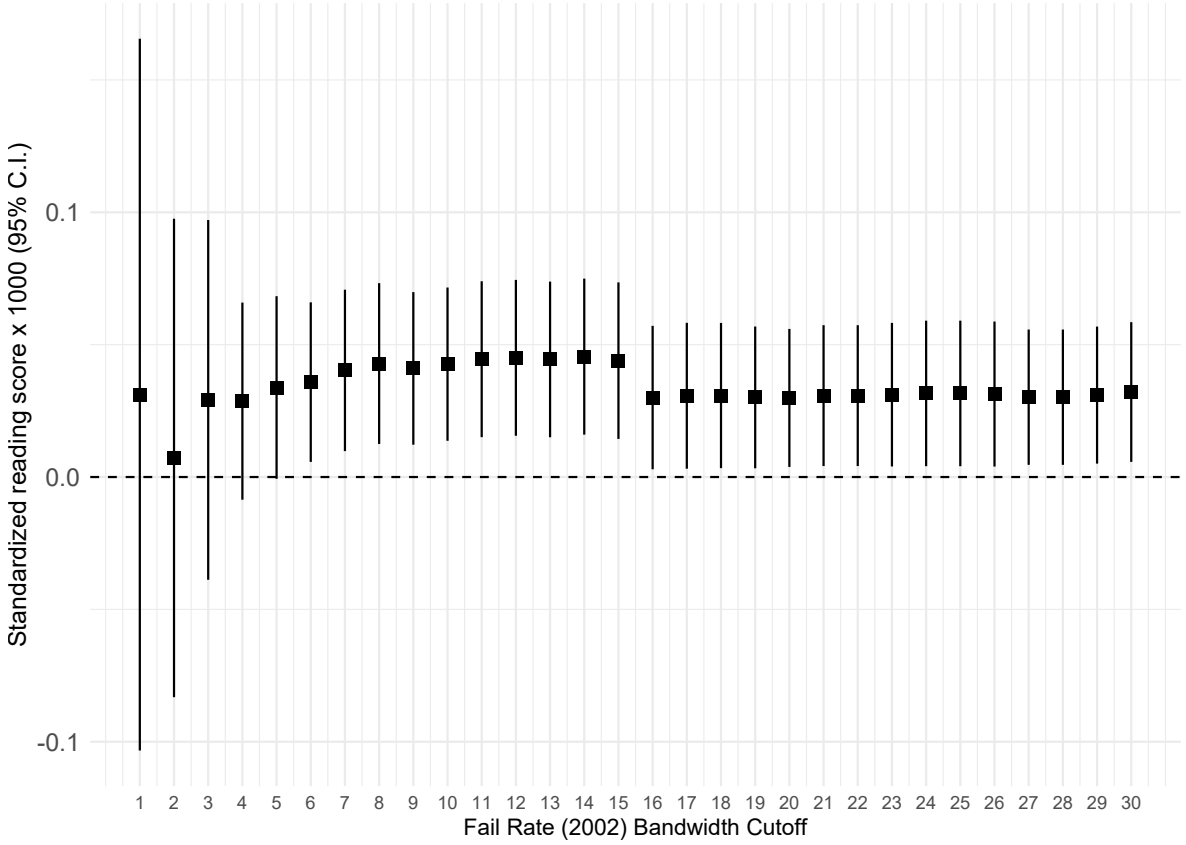
Notes: Same specification as the main reading persistence figure, but restricted to students who attended a non-RF district through third grade and moved into an RF district beginning in fourth grade. Each horizontal panel reports pooled difference-in-differences estimates for progressively longer windows of student observation. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A6: Effect of Reading First eligibility on third-grade math scores



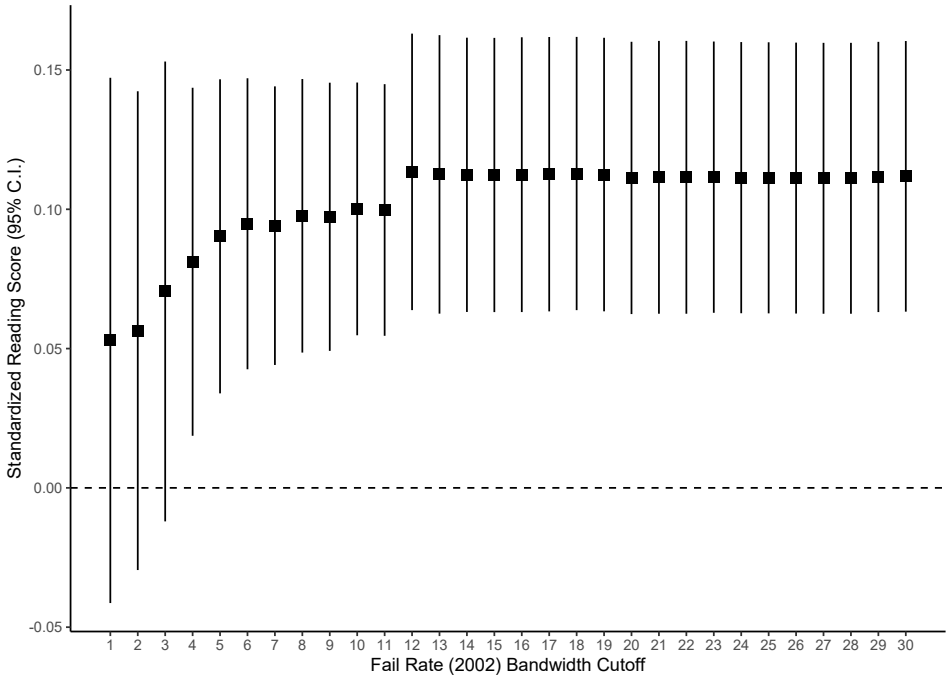
Notes: Event-study estimates of the effect of RF eligibility on standardized third-grade math scores, from the same specification as the main reading event study (equation 2). All models include campus-by-grade and year fixed effects and demographic controls (share Black, Hispanic, Asian/Native in 2002) interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A7: Effect of Reading First expenditures on third-grade reading scores, varying bandwidths around 2002 fail rate of 13.6, IV specifications



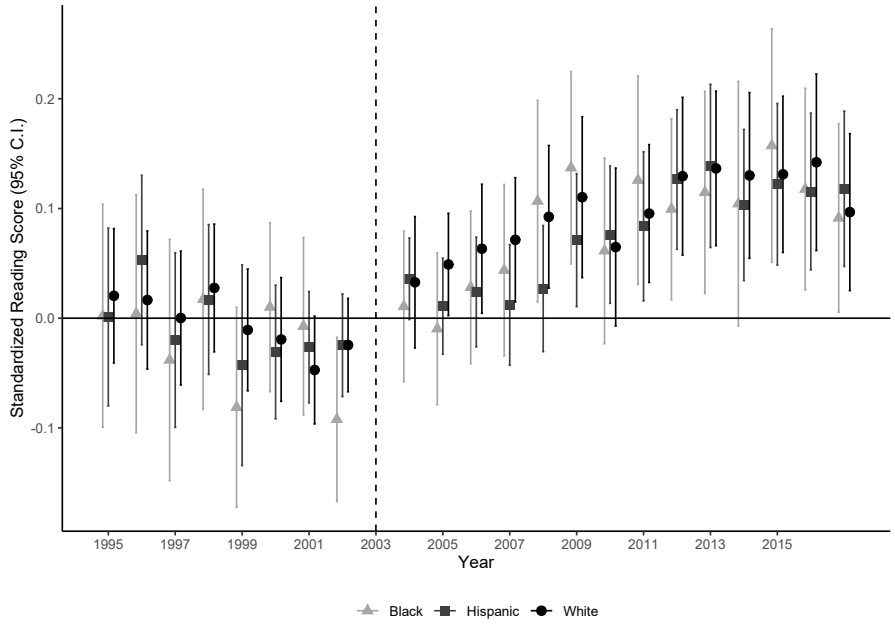
Notes: Each point reports the 2SLS estimate of cumulative RF funding per K-3 students on standardized third-grade reading scores, restricting the sample to districts within the indicated bandwidth (in percentage points) of the 13.6% fail rate eligibility threshold. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A8: Effect of Reading First eligibility on third-grade reading scores, varying bandwidths around 2002 fail rate of 13.6, DiD specifications

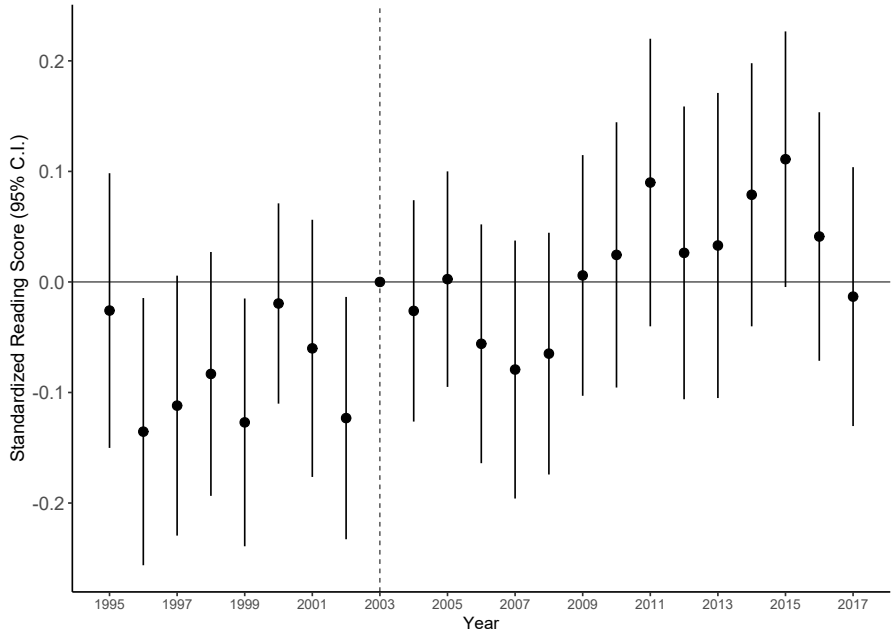


Notes: Each point reports the pooled difference-in-differences estimate of the effect of RF eligibility on standardized third-grade reading scores, restricting the sample to districts within the indicated bandwidth (in percentage points) of the 13.6% fail rate eligibility threshold. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A9: Effect of Reading First eligibility on third-grade reading scores, by race



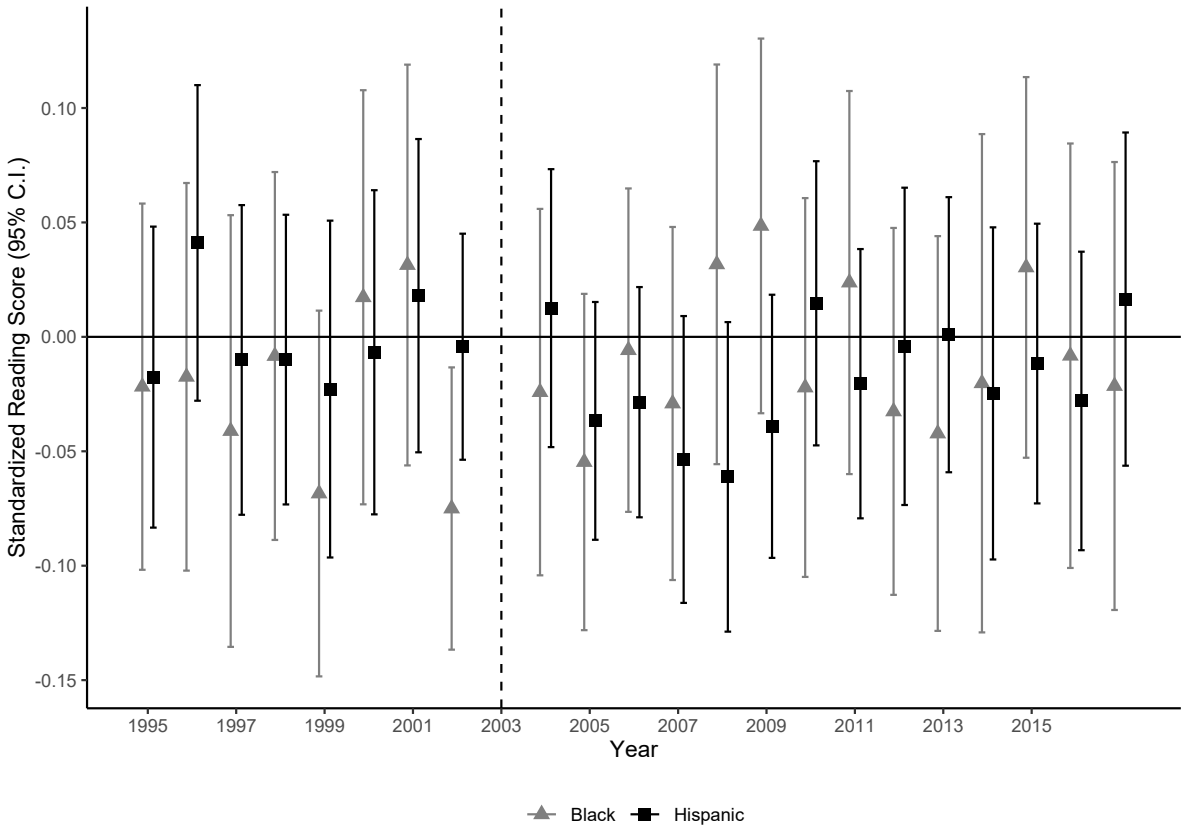
(a) Black, White, and Hispanic



(b) Asian and Native American

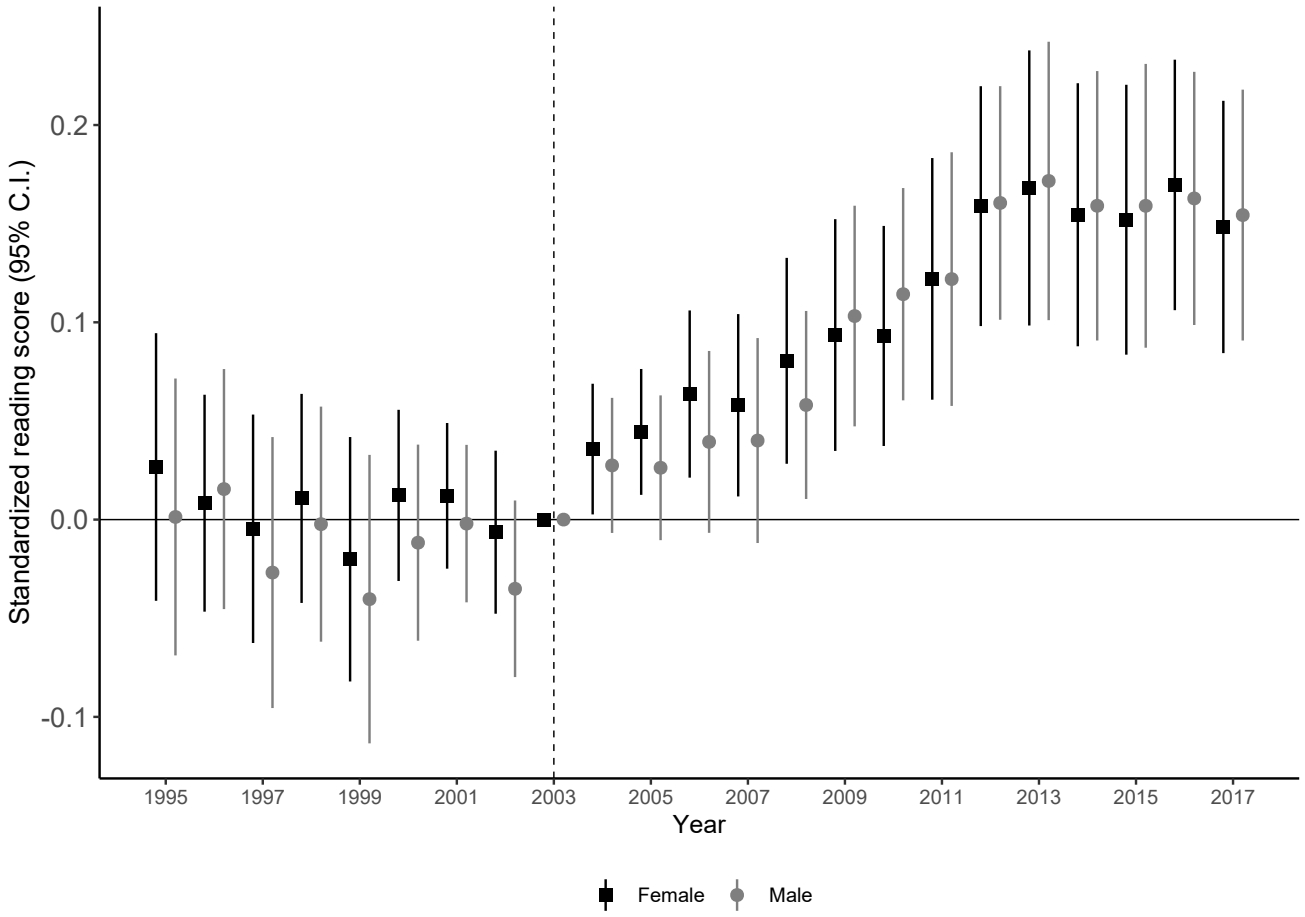
Notes: Event-study estimates of the effect of RF eligibility on standardized third-grade reading scores, estimated separately by race. Panel (a) shows results for Black, White, and Hispanic students. Panel (b) shows results for Asian and Native American students. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A10: Effect of Reading First eligibility on third-grade reading scores, Hispanic–White and Black–White race gap



Notes: Event-study estimates where the outcome is the difference in standardized third-grade reading scores between Hispanic and White students (and between Black and White students). All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

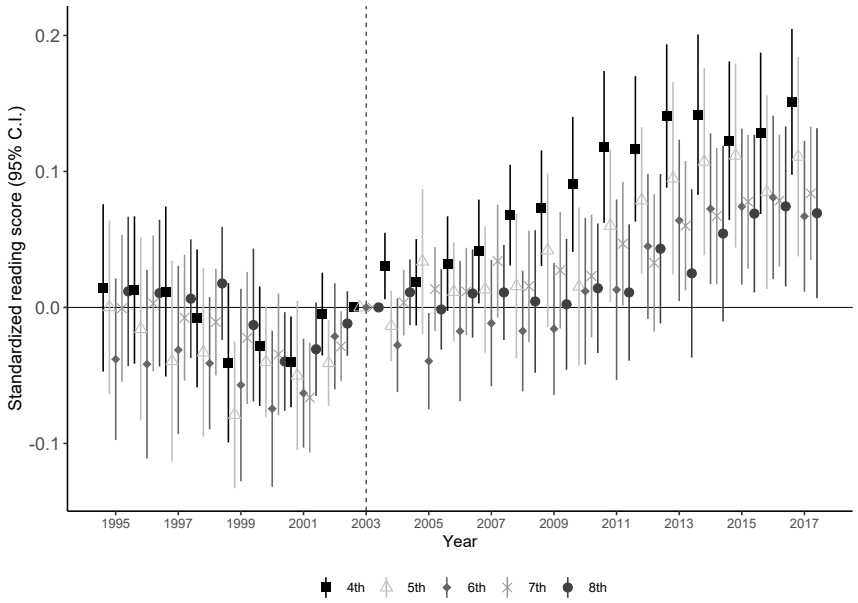
Figure A11: Effect of Reading First eligibility on third-grade reading scores for males and females



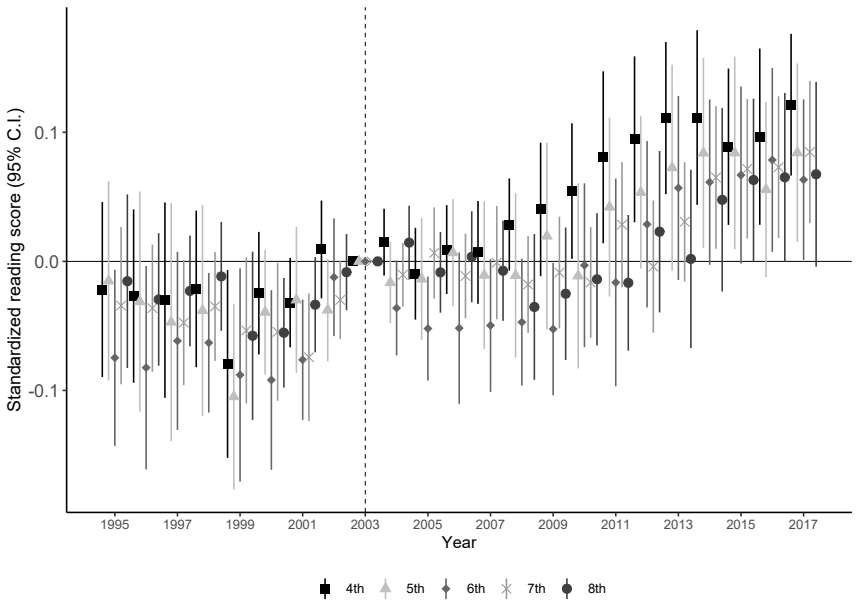
Notes: Event-study estimates of the effect of RF eligibility on standardized third-grade reading scores, estimated separately for male and female students. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A12: Effect of Reading First eligibility on reading scores, grades 4–8

(a) All

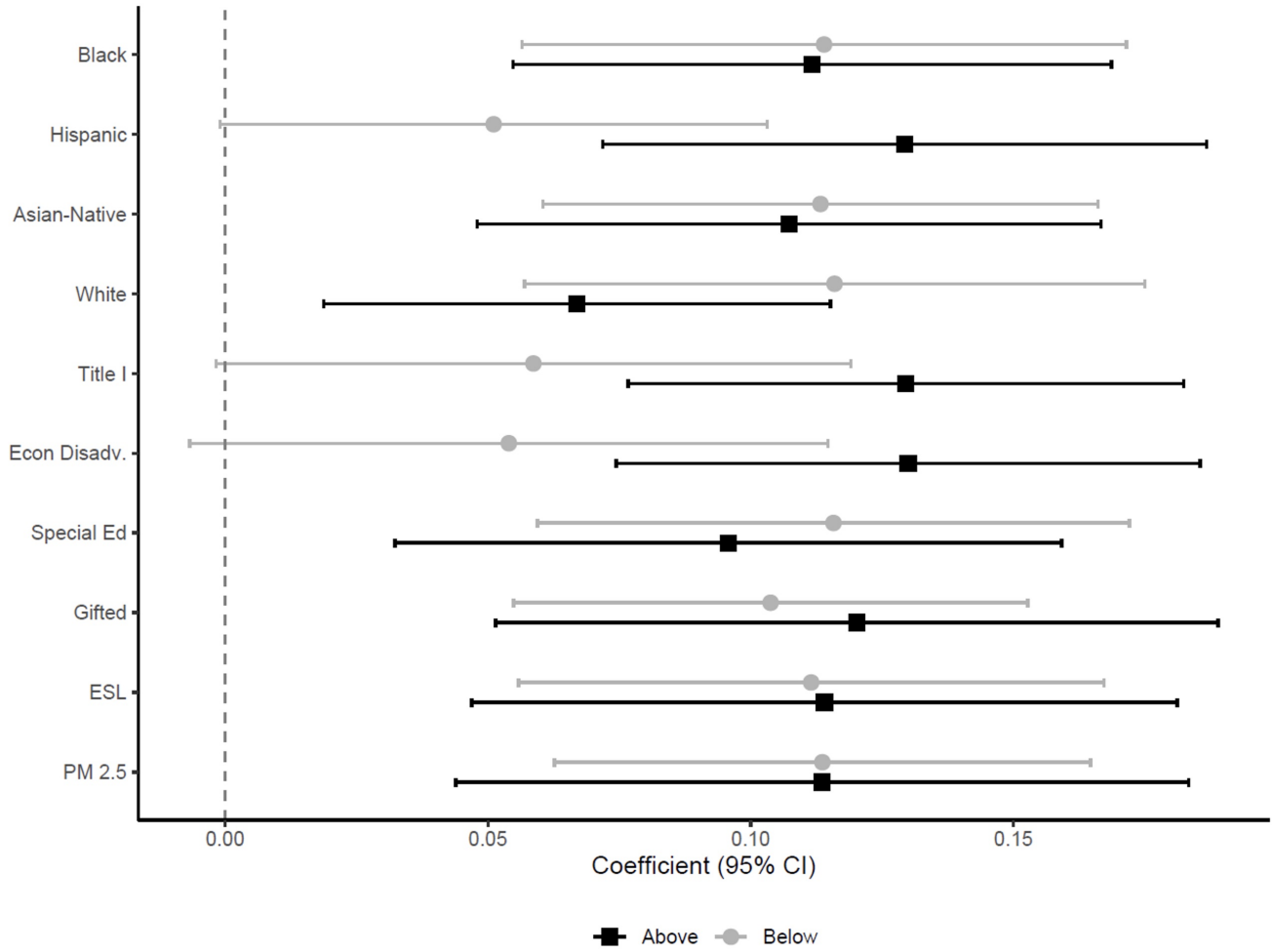


(b) Restrict to 2002 fail rate ≤ 5



Notes: Event-study estimates of the effect of RF eligibility on standardized reading scores for grades 4–8. Panel (a) includes all districts; panel (b) restricts to districts within five percentage points of the 13.6% fail rate eligibility threshold. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

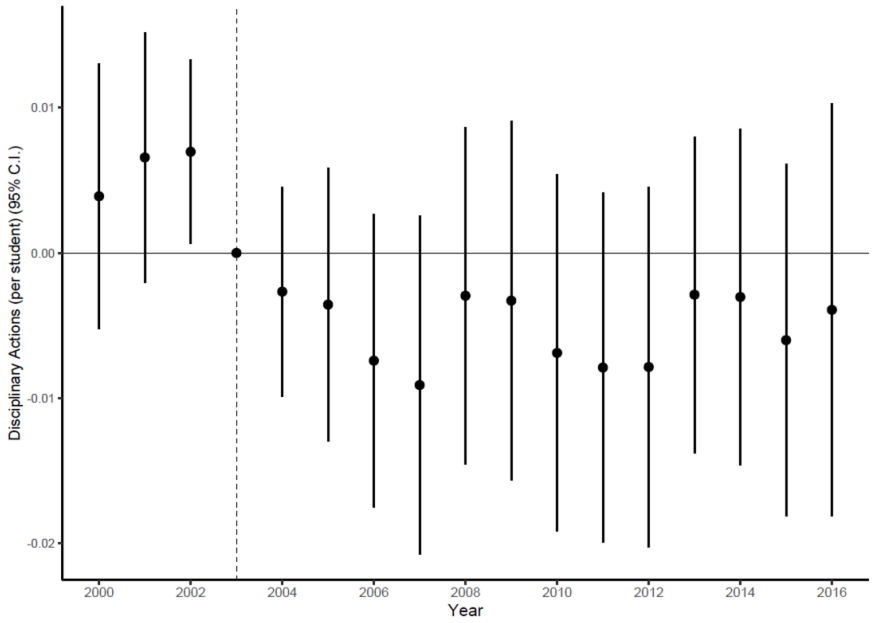
Figure A13: Heterogeneity analysis of the effect of Reading First eligibility on third-grade reading scores



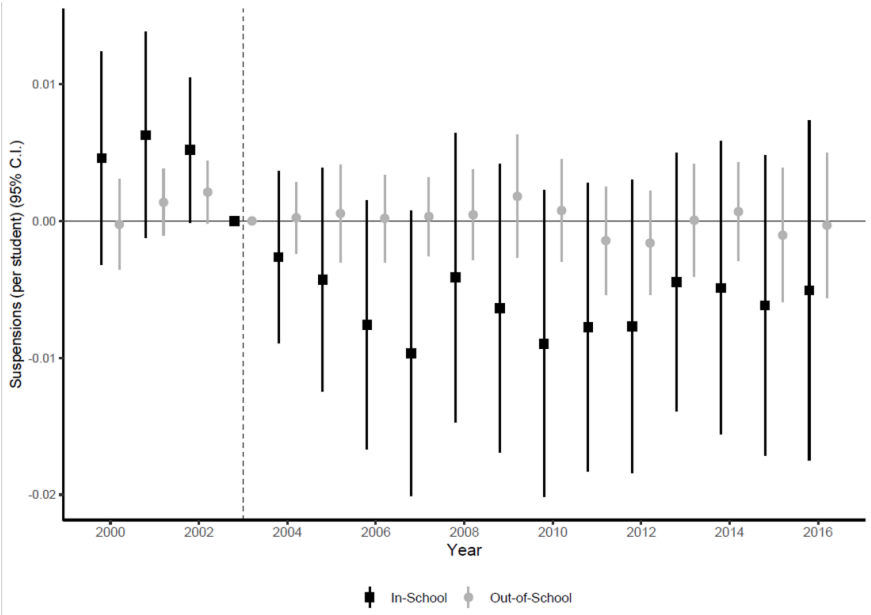
Notes: Each row reports split-sample difference-in-differences estimates, where campuses are divided at the median of the indicated characteristic measured in 2002. The sample is split into above and below median based on a series of campus observables in the year 2002. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A14: Event study of Reading First eligibility on third-grade disciplinary outcomes

(a) All disciplinary outcomes

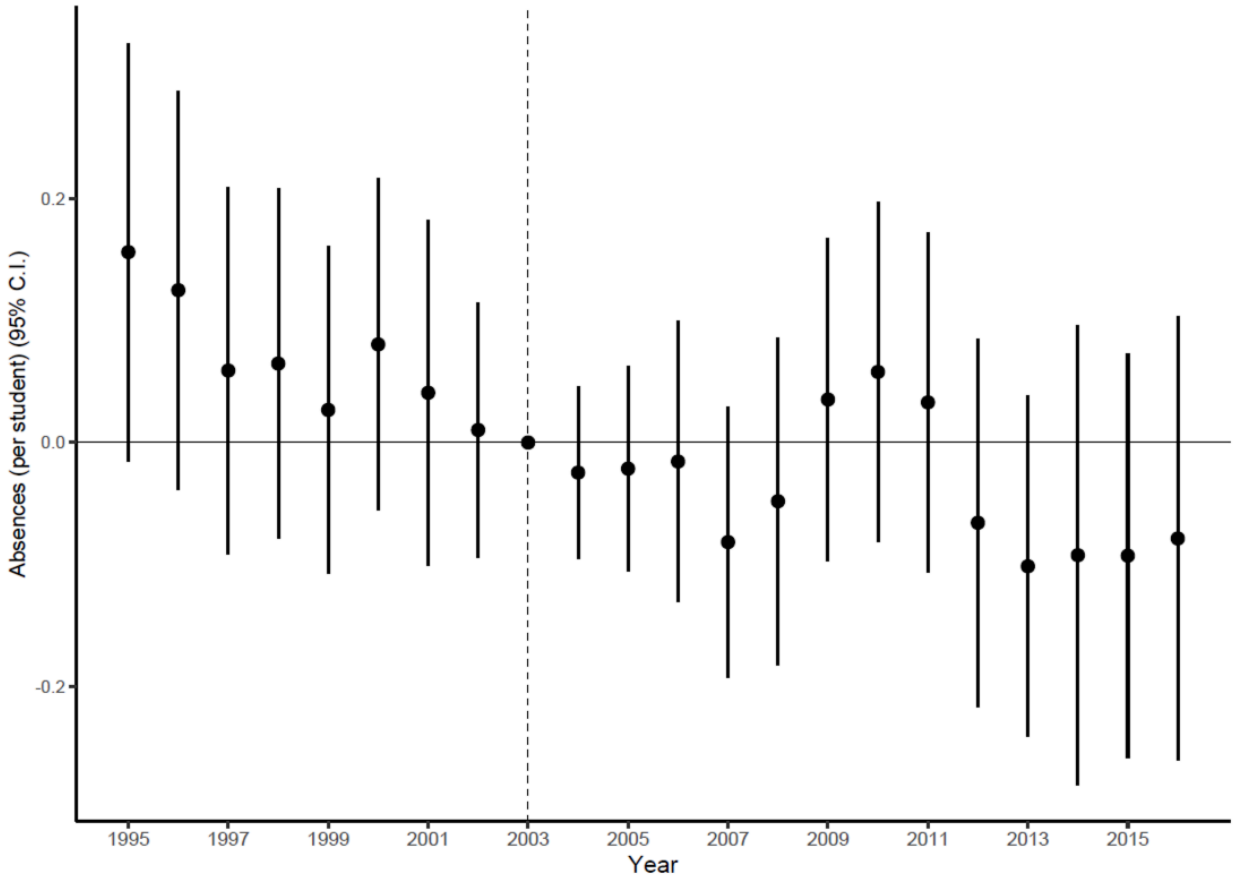


(b) In-school suspensions (restrict 2002 fail rate ≤ 5)



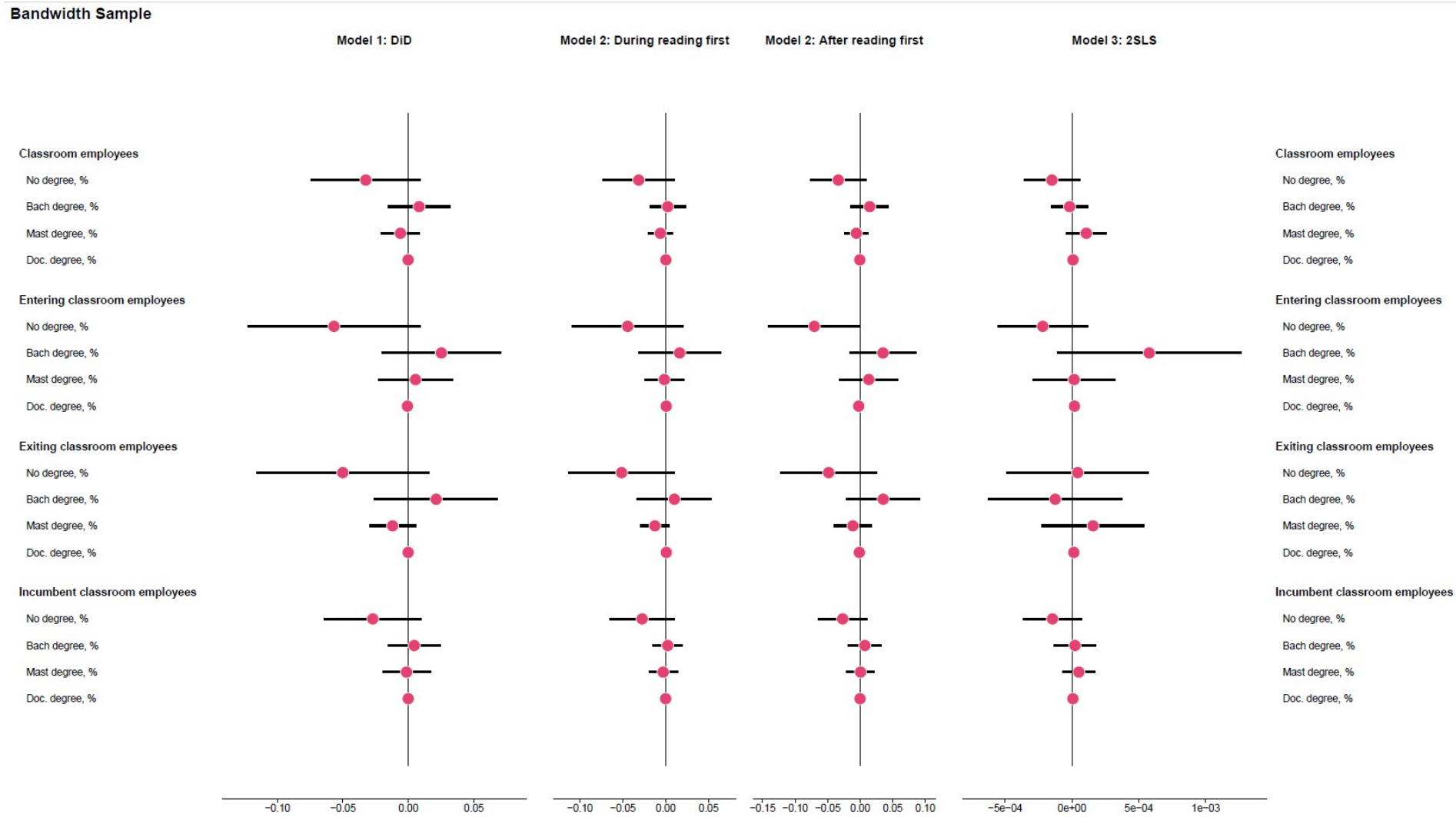
Notes: Event-study estimates of the effect of RF eligibility on third-grade disciplinary outcomes per student. Panel (a) includes all disciplinary actions (in- and out-of-school suspensions and expulsions) for all districts. Panel (b) shows in-school suspensions only, restricted to districts within five percentage points of the 13.6% fail rate eligibility threshold. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A15: Effect of Reading First eligibility on third-grade absences



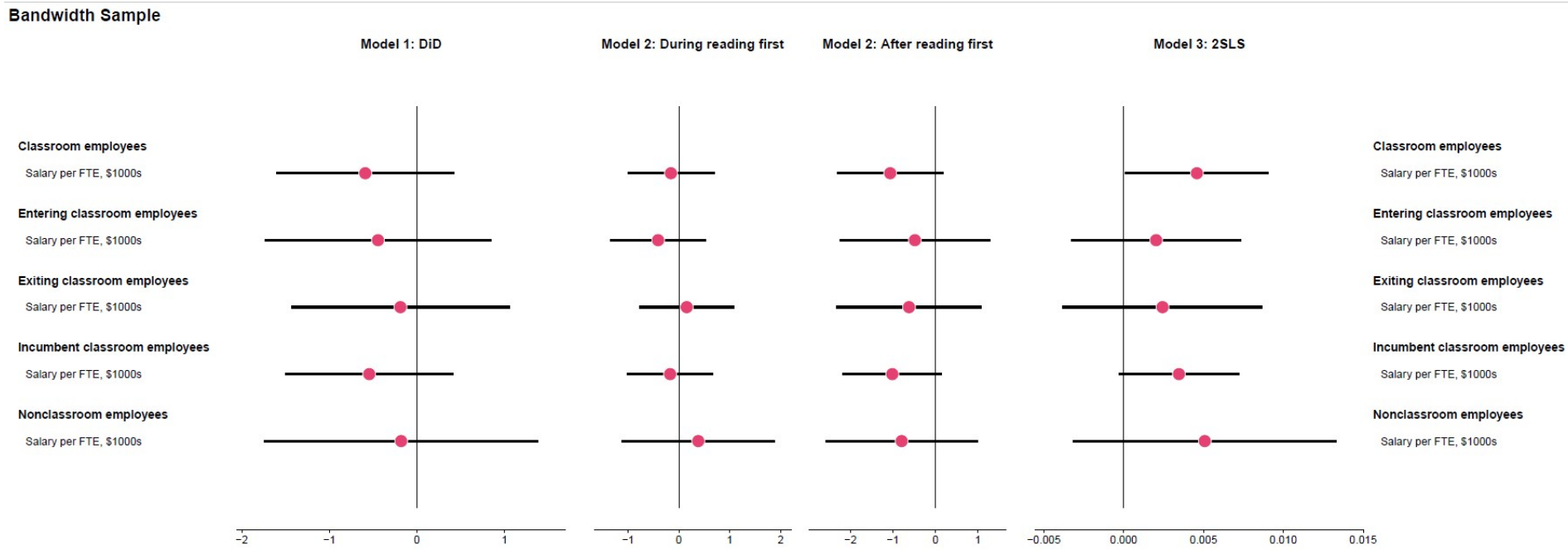
Notes: Event-study estimates of the effect of RF eligibility on campus-grade absences per student. All models include campus-by-grade and year fixed effects and demographic controls interacted with year indicators. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A16: Changes in teacher degree attainment



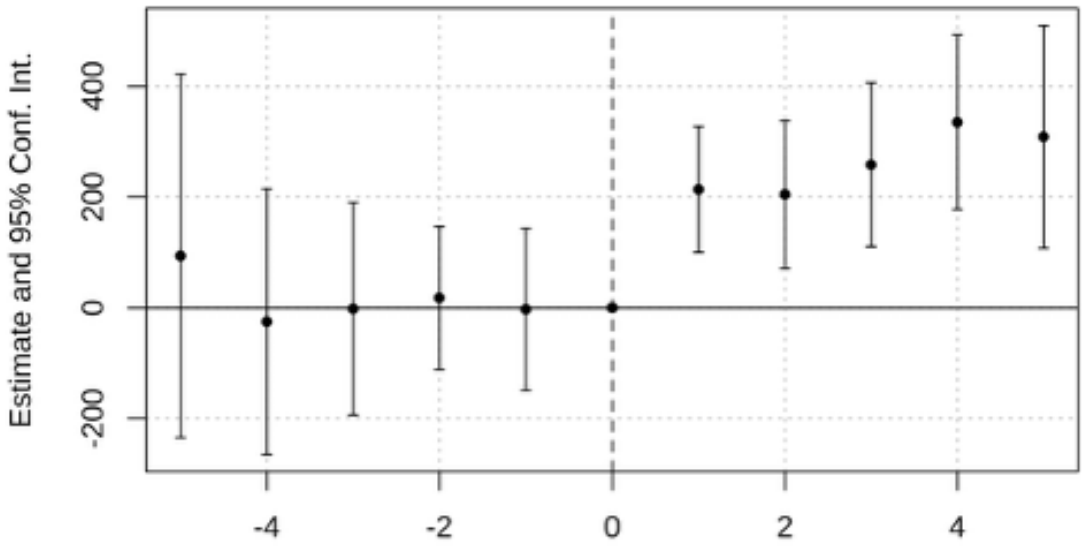
Notes: Each panel reports difference-in-differences, separate-period, and 2SLS estimates of the effect of RF eligibility on the share of campus teachers with various degree levels. Results are shown for all classroom employees, entering teachers, exiting teachers, and incumbents. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A17: Changes in teacher salaries



Notes: Each panel reports difference-in-differences, separate-period, and 2SLS estimates of the effect of RF eligibility on teacher salary per FTE (in thousands of dollars). Results are shown separately for all classroom employees, entering teachers, exiting teachers, incumbents, and nonclassroom employees. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A18: Effect of tax election on local tax revenue



Notes: Stacked difference-in-differences event-study estimates of the effect of passing a school district tax election on total per-pupil local revenue. Treated districts are those where a tax increase passed; the comparison group consists of districts where a similar measure was on the ballot but failed. All specifications include district and year fixed effects, weighted by total enrollment. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Table B1: Effect of Reading First eligibility on Reading First school district revenue, per total students

	DiD		DiD, separate periods		First-stage	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Reading first revenue per student						
1(RF Eligible) × 1(year>2003)	15.59***	13.20***				
	(5.73)	(3.02)				
1(RF Eligible) × 1(2004≤year≤2011)			31.36***	27.13***		
			(4.73)	(6.03)		
1(RF Eligible) × 1(year>2011)			0.60	0.13		
			(0.56)	(0.39)		
Predicted RF $ per student					0.21***	0.31***
					(0.04)	(0.08)
F-stat (instrument)					31.72	16.14
Observations	24,121	8,630	24,121	8,630	24,121	8,630
District FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Weights	All students	All students	All students	All students	All students	All students
Bandwidth	None	5	None	5	None	5

Notes: Each column reports a separate regression of RF-specific revenue per total student 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 periods. Columns (5)–(6): first-stage regression of actual RF revenue on predicted RF revenue per total student. 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 enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

	DiD		DiD, separate periods		First-stage	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Non-reading first revenue per student						
1(RF Eligible) × 1(year > 2003)	163.71	145.24				
	(159.50)	(178.86)				
1(RF Eligible) × 1(2004 ≤ year ≤ 2011)			179.48	188.01		
			(129.86)	(143.74)		
1(RF Eligible) × 1(year > 2011)			148.71	105.14		
			(176.97)	(225.87)		
Predicted RF $ per student					−0.04	0.23
					(0.33)	(0.87)
F-stat (instrument)					0.01	0.06
Observations	24,121	8,630	24,121	8,630	24,121	8,630
District FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Weights	All students	All students	All students	All students	All students	All students
Bandwidth	None	5	None	5	None	5

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 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 periods. Columns (5)–(6): first-stage regression of non-RF revenue on predicted RF revenue per total student. 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 enrollment. Standard errors clustered at the district level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: Effect of Passing a Tax Election on District Finances

	TWFE	2SLS IV
	(1)	(2)
Panel A: Revenue		
Local Revenue	312.557***	
	(93.920)	
State Revenue	-41.407	
	(70.534)	
Federal Revenue	-9.087	
	(35.408)	
State + Federal Revenue	-49.915	
	(85.814)	
Panel B: Broad Expenditure		
Total School Expenditure	183.046**	0.586**
	(72.738)	(0.250)
Salary, Wages & Benefits	158.211***	0.507**
	(52.130)	(0.233)
Travel	1.087	0.003
	(1.794)	(0.006)
Professional Services	21.976	0.070
	(43.867)	(0.130)
Supplies & Materials	16.944	0.054
	(15.384)	(0.054)
Misc. Operating Costs	-15.398	-0.049
	(10.812)	(0.040)
Panel C: Specific Expenditure		
Supplies & Materials	16.944	0.054
	(15.384)	(0.054)
Books (Total)	0.342	0.001
	(10.465)	(0.030)
General Supplies	19.804	0.063
	(14.052)	(0.051)
Testing Materials	-1.639***	-0.005**
	(0.573)	(0.002)
Panel D: Books		
Books (Total)	0.342	0.001
	(10.465)	(0.030)
Library Books	1.518	0.005
	(2.167)	(0.007)
Textbooks	-1.216	-0.004
	(3.531)	(0.011)
Reading Materials	-4.818**	-0.016*
	(2.453)	(0.009)
Observations	16,459	16,459
District & Year FE	Yes	Yes
Weights	Enrollment	Enrollment
Clustering	District	District

Notes: Each row reports the coefficient from a separate regression. Column (1) reports TWFE estimates of the effect of passing a tax election. Column (2) reports 2SLS estimates using the tax election passage as an instrument for local revenue. All specifications include district and year fixed effects, weighted by total enrollment, with standard errors clustered at the district level. Panel A reports revenue outcomes (IV not applicable as local revenue is the endogenous variable). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.