Decoding Success: Phonics-based Instruction and Student Achievement*

Alex Hollingsworth, Alberto Ortega, and Nikolaos Zirogiannis†

November 9, 2026

Abstract

Reading is a foundational skill, yet fewer than one in three US fourth graders read at a proficient level. How best to teach reading and improve educational attainment has been the subject of an active policy debate for decades, with some arguing for unrestricted funding increases and others supporting targeted changes to specific pedagogy. Recently, phonics-based approaches have surged in popularity, even though the largest attempt to put similar methods into classrooms was previously abandoned out of concerns that the program was ineffective. This paper studies Reading First—a nearly $1 billion per year federal program established under No Child Left Behind that promoted phonics-based instruction in K–3 classrooms. Exploiting administrative data from Texas and program eligibility formulas, we estimate effects via intent-to-treat difference-in-differences and instrumental variables specifications. We find that eligibility increased third-grade reading test scores by 0.1 standard deviations, and effects persisted through at least the eighth grade. An analysis of digitized exam questions shows that Reading First students outperformed peers on questions that required basic decoding skills and those with more words, a benefit that also drove improvements in math scores. District expenditure data show that Reading First dollars were spent on reading materials and teacher salaries. Finally, we follow exposed cohorts into early adulthood finding that the phonics-based intervention increased on-time high school graduation, college enrollment, four-year college completion, and adult earnings at age 25 by around $500.

JEL Classifications: H75, I21, I22, I28
Keywords: Literacy, Phonics, Reading First, School Spending

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

†Hollingsworth: Ohio State University Department of Agricultural, Environmental, and Development Economics; Department of Economics; and NBER (hollingsworth.126@osu.edu); Ortega: Indiana University – Bloomington O'Neill School of Public and Environmental Affairs, and NBER; (alorte@iu.edu), and Zirogiannis: Indiana University – Bloomington O'Neill School of Public and Environmental Affairs (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.

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

Society has many levers to improve literacy. Within the classroom, research has highlighted the critical roles of teacher quality and class size, among other inputs (Hoxby, 1996; Krueger, 1999; Chetty et al., 2014b).[1] Moreover, additional school funding has been shown to raise achievement on average (Jackson et al., 2016; Jackson and Mackevicius, 2024), but there is less agreement over which specific investments produce gains or whether districts do better with unrestricted funds (Handel and Hanushek, 2023). Despite this large literature, the specific pedagogical methods used in the classroom are notoriously difficult to observe in administrative data and thus have received relatively less attention. Curriculum and instructional materials have been characterized as the "forgotten" inputs in the education production function, frequently overshadowed by more expensive structural reforms (Whitehurst, 2009; Boser et al., 2015).

American classrooms have long been divided in how reading is taught. Under phonics-based instruction, children learn to map letters to sounds. In contrast, "whole-language" approaches emphasize context clues and text immersion (Goodman, 1967; Clay, 1991). The divide between these two schools of thought grew entrenched enough to become known as the "reading wars" (Mathews, 1966; Pearson, 1989; Kogan, 2025). Decades of research, ranging from laboratory experiments to brain imaging, has shown that children learn to read most effectively through phonics (National Reading Panel, 2000; Rayner, 1998; Castles et al., 2018). However, the largest previous attempt to put these methods into classrooms was abandoned out of concerns that the program was ineffective (Gamse et al., 2008). Nonetheless, a recent "science of reading" movement reviving the phonics approach has surged in popularity, with over 40 states enacting legislation mandating phonics-based instruction in early literacy curricula (Schwartz, 2022; Kogan, 2025; Hanford, 2020). States are thus rapidly adopting a pedagogy that is proven in the laboratory but unproven as a broad policy lever.

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

[1] Numerous studies also examine the influence of factors external to the classroom such as socioeconomic status, neighborhood violence, segregation, pollution exposure, nutrition, and the home language environment (Reber, 2010; Dahl and Lochner, 2012; Anderson et al., 2018; Ang, 2021; Persico and Venator, 2021; Cabral et al., 2021; Hollingsworth et al., 2022; Cox et al., 2026; Andersen et al., 2025; Ortega and Ludwig, 2022).

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

To identify the causal effect of Reading First, we exploit the quasi-experimental variation created by funding rules in Texas. Program eligibility was determined based on preexisting test scores and district poverty indicators. Among eligible districts, the amount of support received was determined by a funding allocation formula. We leverage this structure to apply two complementary identification strategies: an intent-to-treat difference-in-differences approach that compares eligible and ineligible districts before and after the program's start and an instrumental variables framework that uses a revenue formula as a simulated instrument for district-level program expenditures. Our analysis uses 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 increased third-grade reading test scores by approximately 0.1 standard deviations relative to the scores in ineligible districts. The effects grew during the funding period, and persisted at a positive and statistically significant level even after the last Reading First dollar was spent. Students exposed to the program in early grades continued to relatively outperform their unexposed peers on the state's standardized reading exam through eighth grade. We also find positive effects on third-grade math scores, driven by questions with more words, suggesting broad improvements in reading skills, not simply that treated schools began focusing more on reading exam performance.

Our results are robust. Event-study estimates show no pretrend differences between eligible and ineligible districts. We find similar effects across a battery of specification checks featuring alternative controls and fixed-effects, dropping controls, unbalanced panels, varying bandwidths for inclusion in the eligibility formula, clustering at different levels, and including flexible time trends by school poverty status. Robustness tests where we restrict comparisons to be only among similar schools (e.g., only the poorest schools) or similar students (e.g., only the poorest students) show that our findings are not driven by targeting or by a particular type of school or student. We find positive and statistically significant effects across virtually every within-group comparison, although, we do find some evidence that the effects were largest in schools with above-median shares of economically disadvantaged, Hispanic, and Title I students. Finally, our instrumental variables approach shows similar results, indicating that districts receiving more revenue experienced larger gains, with the marginal benefit of the treatment from additional dollars decreasing with spending.

We follow the earliest Reading First–exposed cohorts as they aged, first through high school

and then into their mid-20s, and find that the short-run achievement gains translated into lasting improvements in human capital. Eligible students are more likely to have graduated from high school on time, have enrolled in college by age 19, and have completed a four-year college degree, with a small offsetting decline in their likelihood of attaining a two-year degree. We also find that eligibility raised earnings at age 25 by $509, with most of that gain running through improved college enrollment and occupational upgrading.

The program was cost-effective and paid for itself. Eligible districts received roughly $75 per K–3 student per year (in 2025 $) over the post period (about $132 per student–year during the program's funding period), and the program produced a pooled improvement of approximately 0.1 standard deviations in reading test scores. Even attributing the entire during-program cost of $132 per student–year to the smaller during-program effect of 0.07 standard deviations, the implied cost per 0.1 standard deviations is $184 per student–year, a fraction of the $3,846 (in 2025 $) per pupil per year that Jackson and Mackevicius (2024) estimate is typically required for a comparable gain. Under the Marginal Value of Public Funds framework of Hendren and Sprung-Keyser (2020), the tax revenue recovered from the higher earnings of treated students is projected to cover the program's full cost by the time treated students reach the age of 31.

We explore several channels through which the program may have improved achievement. First, we collect, digitize, and classify the content of third-grade reading and math exams. We then link these to individual student answers, finding that Reading First students relatively outperformed their untreated peers on questions with more words, questions requiring broad comprehension of a passage, and questions whose words are easier to decode. The same advantage appears on math exams, where Reading First benefits are larger on questions with more and easier to decode words. Treated students show no detectable benefit for math questions that are closest to pure computation. These findings are consistent with the hypothesis that phonics instruction improves decoding skills, which in turn helps students engage with more demanding text. Second, using detailed district expenditure data, we show that Reading First funds were spent primarily on reading materials, general instructional supplies (including materials such as workbooks), and staff wages. Third, we find little evidence that schools in Reading First districts attracted more or higher-quality teachers. Finally, we compare Reading First expenditures to a more general revenue shock, the passage of school district tax elections. Districts spent election revenue mainly on salaries, and the gain in reading scores per dollar was smaller than under Reading First. The contrast suggests that the curricular requirements attached to Reading First dollars contributed to their effect beyond the resources themselves.

Our work makes several contributions to the literature. First, we provide causal estimates of Reading First's effect on student outcomes. Despite being the academic cornerstone of No Child Left Behind and costing over $1 billion per year nationally, the program has received little attention in the economics literature. Prior work commissioned by the federal government found no measurable effect on student reading comprehension (Gamse et al., 2008). This null finding contributed to Reading First's defunding and shaped a decade of federal education policy. However, the Gamse et al. (2008) evaluation used data for years prior to 2008 only,

the year the report was published. Because Reading First was a K–3 intervention and the first funding was released in mid-2003, approximately only half 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, consider longer-run effects, and have a larger number of treated schools and students. This provides a more accurate and comprehensive assessment of Reading First's true efficacy.

Second, we present evidence on whether the scientific consensus on phonics instruction can be implemented at scale through government policy. A growing literature examines recent state and international reading reforms. The best known is Mississippi's 2013 literacy law, after which fourth-grade reading scores on the National Assessment of Educational Progress rose, gains that became commonly referred to as the "Mississippi miracle" (Spencer, 2024), although Wainer et al. (2026) argue that accompanying retention policies complicate causal attribution. Reading scores also rose in low-performing California schools that received an Early Literacy Support Block Grant (Novicoff and Dee, 2026). In England, Machin et al. (2018) find that synthetic phonics instruction raised reading scores at age 5, with effects persisting to age 7, but fading by age 11.[2] 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. This allows us to more cleanly capture the effect of the policy than if we studied the same policy today, since the untreated control districts during our study period are less likely than school districts today to have adopted similar phonics practices independently. Our setting also avoids the confounding effects of the COVID-19 pandemic and the related state policy responses (Davis et al., 2021; Bacher-Hicks et al., 2021; Dee, 2024), which may complicate the interpretation of studies examining more recent reforms (Westall and Cummings, 2023).

Third, we contribute to the literature on the persistence of educational interventions and identify the mechanisms through which Reading First improved outcomes. A well-documented pattern is that early interventions produce test score gains that dissipate within a few years, yet some of these same programs still improve long-run outcomes, creating a puzzle about what mediates the long-run effects (Chetty et al., 2011; Jacob et al., 2010; Bailey et al., 2020; Cascio and Staiger, 2012; Gray-Lobe et al., 2023; Banerjee et al., 2007). Our findings diverge from this pattern. Reading First induced gains that persisted through every year we have data to examine and affect long-run outcomes. This provides evidence that learning to read at a young age may generate a "fork in the road" in human capital accumulation (Hoxby, 2021) and is consistent with models of dynamic skill formation where early skills generate compounding returns to future learning (Cunha and Heckman, 2007; Cunha et al., 2010; García et al., 2023).

[2]Westall and Cummings (2023) find gains only where states adopted a full package of early literacy policies and descriptive evidence has shown test score gains in Michigan after its 2016 "Read by Grade 3" law (Strunk et al., 2021). Internationally, remedial reading tutoring in Colombia (Álvarez Marinelli et al., 2024) and a structured early-grade literacy program in Uganda (Buhl-Wiggers et al., 2024) produced gains that persisted for about five years after each program ended, and a structured Arabic literacy curriculum in the United Arab Emirates raised first graders' word reading by 0.19 standard deviations in a randomized trial (John et al., 2026).

Fourth, our work speaks to the economics of federal intervention in a decentralized education system and the design of intergovernmental grants (Oates, 1999). The United States contains thousands of school districts with long traditions of local financing and local governance (Hoxby, 2000; Alesina et al., 2004). Despite this, the federal government has increasingly used conditional grants to influence local educational practices. This design question has received considerable attention in various policy domains (Baicker and Staiger, 2005; Knight, 2002), with more limited evidence in education (Gordon, 2004; Cascio et al., 2010). Much of this work shows that intergovernmental aid to school districts is often diverted or crowded out by local responses (Clemens et al., 2026; Brunner et al., 2020). There is also limited evidence on the effectiveness of standards and mandates that arrive without new resources (Loveless, 2021; Polikoff, 2021). We show that Reading First was a conditional grant that worked as designed. Reading First dollars reached eligible districts, were spent on the practices the program required instead of crowding out other revenue, and produced lasting achievement gains. In a similar vein, our findings add to recent work showing that curriculum content shapes students' knowledge, beliefs, and later choices (Arold, 2024; Ainsworth et al., 2025; Ellison and Pathak, 2025).

Fifth, we ask whether the effects of funding tied to specific practices differ from those of an equivalent unconditional revenue increase. A large literature shows that school spending improves educational and economic outcomes (Jackson et al., 2016; Lafortune et al., 2018; Abott et al., 2020; Baron, 2022; Biasi, 2023; Jackson and Mackevicius, 2024; Kreisman and Steinberg, 2019; Duque, 2026), and recent work has begun to ask which uses of money matter most (Brunner et al., 2020; Jackson and Mackevicius, 2024; Biasi et al., 2025). We contribute to this second question by comparing the effects of Reading First expenditures with those of untargeted increases in district revenue generated by successful local tax elections in the same state and years. Districts spent election revenue primarily on salaries rather than on reading materials or instructional supplies, and the resulting gains in reading scores were smaller per dollar than those of Reading First.

The remainder of the paper is organized as follows. Section 1 provides background on phonics, reading curricula debates, and the Reading First program at both the national and Texas levels. Section 2 describes our data sources, including the administrative education records, district expenditure data, digitized exam questions, and school district tax elections. Section 3 outlines our empirical approach. Section 4 presents our main results on reading achievement along with internal validity checks and an analysis of campus- and student-level heterogeneity. Section 5 explores mechanisms beginning with the question decoding analysis (where we also present the math results), expenditure patterns, heterogeneity in effects by how districts allocated their awards, school and teacher characteristics, and a comparison of Reading First expenditures to school funding from other sources. Section 6 presents long-run results, following exposed students through grades 4–8, high school, college, and into the labor market. Section 7 presents Marginal Value of Public Funds estimates for Reading First, and Section 8 concludes.

1 Background

1.1 Phonics and the Reading Wars

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

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

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

By the early 1960s, phonics was a popular method of reading instruction in the United States in part because of the success of the 1955 book *Why Johnny Can't Read* and Cold War anxieties about American educational standards. However, phonics' foothold was soon replaced as methods that prioritized context became more prominent in the 1970s. The most popular of these was the whole language approach, whose central tenet is that children can learn to read similarly to how they learn to speak—through repeated exposure to and immersion in the type of text people actually read rather than through the incremental letter–sound pairing skill formation emphasized in phonics. In 1989, P. David Pearson, prominent reading scholar and dean of the Graduate School of Education at the University of California – Berkeley, wrote that he had never "witnessed anything like the rapid spread of the whole-language movement. Pick your metaphor—an epidemic, wildfire, manna from heaven—whole language has spread so rapidly throughout North America that it is a fact of life in literacy curriculum and research" (Pearson, 1989 quoted in Kogan, 2025).

The debate took on a partisan dimension, with phonics increasingly identified with the Republican Party and whole language with Democrats. By the late 1990s, more than 65% of

[3] See Kogan (2025) for an account of this debate and its political dimensions.

[4] The Center for the Improvement of Early Reading Achievement (CIERA) created a simple way to estimate how decodable words are, given the letter and sound patterns in each word (Hoffman et al., 2002; Menon and Hiebert, 1999). The scores range from 1 to 8. The word *cat* has a decodability score of 2 since it is a consonant–vowel–consonant pattern that does not end in an *r* or an *l*; *coat*, with the vowel letter pair, has a score of 5; and the word *caught* would be the most difficult to decode, with a score of 8. We use the CIERA decodability scorer to measure the share of words within each exam question that can be decoded using basic phonics skills in Section 2.2.

state legislators introducing phonics bills were Republicans (Paterson, 2000). In 1997, Congress instructed the National Institute of Child Health and Human Development to convene an expert panel to synthesize the evidence on reading instruction. The resulting National Reading Panel (2000) endorsed systematic and explicit phonics instruction, along with emphasis on phonemic awareness, fluency, vocabulary, and comprehension.

1.2 National Reading First Initiative

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

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

Reading First was administered as competitive grants to states. Each state agency submitted a detailed proposal to the federal government that outlined how the program would disburse funds and operate. States successful in obtaining a Reading First grant administered their own competitive subgrants to school districts, each of which would need to submit their own application. Subgrants were supposed to be awarded on the basis of need, and state educational agencies were instructed to prioritize districts with lower reading performance and high poverty (McCallion, 2005). In practice, the program reached nearly every state and served thousands of elementary schools nationwide, although exactly how funds were administered to school districts differed by state. Most of the initial Reading First awards were made to states between the summer of 2003 (after the 2002–2003 exams were taken) and 2005, with the first fully treated set of students to go through the program taking their third-grade exams in 2007.

No Child Left Behind also mandated an independent national evaluation of Reading First's effectiveness, conducted by the US Department of Education. An initial impact study examining the effects of Reading First between 2004 and 2007 for 17 school districts across 13

states found that the program led to changes in teaching practices. Schools receiving Reading First funds spent more time teaching phonics-based approaches to reading and provided more teacher training and coaching. The report did not find statistically significant gains in student reading comprehension by third grade (Gamse et al., 2008). However, the estimated effects on reading comprehension in grades one through three were positive but small and not statistically significant.

A separate federal implementation study used a descriptive analysis across 24 states comparing outcomes in schools with Reading First funds to those in schools not receiving funding. This report similarly found that Reading First schools devoted more time to reading instruction and adopted the required instructional elements; that nearly 100% of Reading First schools had a reading coach by the end of the program, compared to 57% of non–Reading First schools; and that Reading First schools showed small, but increased reading performance relative to non-Reading first schools through 2006 (Moss et al., 2008). Since the report compared practices at Reading First schools to non-Reading First schools and Reading First was not randomly assigned to schools across the country, the report appropriately cites this as a caveat in the interpretation of its findings. Furthermore, these studies were conducted with data for 2004–2007, a period where the vast majority of students within the study had yet to be exposed to phonics instruction for their entire K–3 school years.

The lack of statistically significant findings in Gamse et al. (2008) prompted questions about whether an intensive focus on phonics and other basic skills was enough to boost reading comprehension. These null findings coupled with concerns about the program's administrative efficiency (McCallion, 2005; Klein, 2008) ultimately led to the end of Reading First.[5] In 2008, Congress cut Reading First's funding by over $600 million (60%) and then eliminated it entirely after fiscal year 2009 (U.S. Department of Education, 2011; Klein, 2008; Robelen, 2011). Although the end of federal appropriations occurred in 2009, previously awarded funds were allowed to be spent by school districts for several more years.

1.3 Reading First in Texas

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

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

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

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

Figure 1A presents a timeline for the rollout 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, with priority given to high-poverty schools with low reading performance. Texas distributed the list of eligible school districts in May of 2003. Texas ran two competitive subgrant cycles for eligible districts (Texas Education Agency, 2003a, 2004). The first cycle was in the summer of 2003 and the second was in the summer of 2004. Once funded, school districts continued participation in the program by applying for continuation awards, which were available until the end of federal funding in 2009. After the second cycle no additional funding cycles were held. Of the 1,128 districts included in our balanced sample, 205 of them ever received Reading First funding, 110 of them first received funding for fiscal year 2004 and 95 of them first received funding afterwards. We discuss the five components of district eligibility and the formula that guided eligible district funding allocation in detail in Section 3 as this is central to our empirical approach.

As indicated in Figure 1B, district receipts and expenditures accordingly ramp up over fiscal years 2004 and 2005. At its peak, the program operated across roughly 700 campuses statewide, with Texas receiving approximately $100 million per year in federal funds. Of every federal dollar received, the state had to distribute at least 80% to school districts. It could keep the remainder 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 Texas. Figure 1C shows that the state abided by this 80% limit as school districts received and spent about 83% of the federal money, slightly above the 80% floor.

The gray solid line in Figure 1D indicates that by 2005, nearly all (98.5%) of the school districts ever awarded a subgrant had received at least some funding. Since K–3 spans four academic years, a student must have entered kindergarten and remained through third grade in a treated district to be considered "fully treated." The dashed line in Figure 1D tracks the share of third graders in treated districts who were fully treated in this sense. The first fully treated cohort entered third grade in 2006 and took exams in 2007. The share of fully treated cohorts continued rising until 2008, when all the third-grade classrooms in treated districts were fully treated since they had been exposed to Reading First curricular interventions for their entire K–3 period. In other words, a district first funded in 2005 would not have produced fully treated third graders until 2008. Figure 1E shows the geographic distribution of school districts that received Reading First funding, by Texas county. We discuss how Texas distributed the money to school districts and what this money was spent on in more detail in Section 2.3.

2 Data

2.1 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), Texas Assessment of Knowledge and Skills (TAKS, 2003–2011), and State of Texas Assessments of Academic Readiness (STAAR, 2012–present). We standardize the scores to have mean zero and standard deviation one among all test takers of a given test, within each state–test–grade–year cell. Thus, our estimates are relative to the statewide score distribution of the exam actually taken that year, and year fixed effects absorb any statewide level shifts across exam regimes. The ERC data also contain rich student demographic information, including race and ethnicity, gender, 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 at grades 4–8 since the influence of exposure to Reading First instruction in early grades may have persisted into later years. The main estimation sample is a balanced panel of campuses observed in every year over the 1995–2017 period. Schools that opened or closed during the study window are excluded from this baseline, though our results are robust to using an unbalanced sample and extending the analysis period through 2019. Table B1 presents descriptive statistics for the full balanced-panel sample of 1,128 districts versus the sample of districts within five percentage points of the state's 2002 average fail rate of 13.6%, one of the five criteria for Reading First funding eligibility, discussed in detail in Section 3.1.

Our baseline estimation sample includes only *first-time* test takers, ensuring that those repeating a grade and retaking a test are not driving our results. We also consider an analysis using all tests taken. 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). These behaviors also capture noncognitive skills that predict long-run outcomes beyond what test scores measure (Jackson, 2018).

To study long-run outcomes, we follow third-grade cohorts through 2014 into high school and cohorts through 2009 into their mid-20s. We then explore whether program participation gains translated into longer-run improvements in high school completion, college enrollment and attendance, and early-career earnings. This approach allows us to trace the effect of Reading First from elementary school all the way to early adulthood. College enrollment and graduation data come from the Texas Higher Education Coordinating Board (THECB) and

the National Student Clearinghouse (NSC), which track enrollment and degree completion at public, community, independent, career, and health-related institutions in Texas. Our labor market results, covering employment and wages, use quarterly earnings records from the Texas Workforce Commission (TWC). The timing of the treatment and subsequent data availability limit our ability to study labor market outcomes over a longer time horizon. The first exam-year containing students who had any exposure to the Reading First program was in 2004. Given the available labor data, which 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 (kindergartners entering in 2004) were approximately 26 years old in 2025. Our labor market estimates reflect this limitation and capture early career outcomes which are smaller in magnitude than peak later in life earnings.

Finally, the Texas ERC 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 can link students to teachers only for years starting in 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 5.3.

2.2 Exam Questions

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

Since phonics instruction is tightly related to decoding of sounds from letters, the composition of exam questions themselves may matter for how Reading First–exposed students performed. A student who has received phonics instruction may perform better on questions that contain words that require phonics skills to decode. An analysis at the question level enables us to test these hypotheses directly, going beyond average treatment effects and exploring whether Reading First improved performance on exactly the type of questions for which phonics skills are the most important.[6]

Although the Texas Education Research Center administrative data contains the multiple choice answers students provide to each question on an exam they do not provide the questions themselves or contents of the answers. Thus, we collect exams across all the years in our sample from four sources: the Texas Education Agency (TEA) website, the Internet Archive (Wayback

[6]Recent evidence also indicates that the content of a question affects both who answers it correctly and what the question can predict about later-life outcomes. Lee and Schaelling (2026) show that the performance effects of topics and words within reading passages can differ by test taker race and gender. Moreno-Medina et al. (2026) link exam items to adult earnings, finding that among all reading questions, those that test basic reading comprehension best predict future earnings.

Machine), the Texas Public Information Request process, and the Texas State Library. 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.[7]

We digitize the available thirteen reading and math exams and classify each question along two dimensions: 1) total number of words in the question-and-answer pair and 2) the share of words in the question that can be decoded with basic phonics skills. We measure the latter on the basis of the CIERA decodability of the words in each reading question, scored from 1 to 8. A score of 1, marks a word that can be sounded out with basic phonics, such as *me*, *we*, and *be* (Hoffman et al., 2002; Menon and Hiebert, 1999). We use this scorer to find the share of words within each question that the CIERA decodability scorer deems are words that should be decodable using basic phonics, those in categories 1–4.

For reading exam questions, we employ two other classifications 1) whether the question asks students to define or identify the meaning of a specific word (a vocabulary item) and 2) whether the question requires students to read an extended passage and demonstrate broad comprehension. For math questions we also consider whether all four answer choices are numeric only, containing no words.

2.3 District Revenues and Expenditures

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

Since Reading First funds are tracked under a dedicated fund code, we are able to observe both the total amount of Reading First funding districts received and how those funds were spent using the function code. We use the object code to decompose Reading First expenditures into payroll, contracted services, supplies and materials, travel, and other categories. Knowing how districts spent awarded money allows us to explore the mechanisms through which Reading First 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. As indicated in Figure 1B, the expenditures and revenues overlap entirely. PEIMS also categorizes supplies and materials into books and general supplies. Books include reading materials, library books, and textbooks. The classifications within these budgetary data reflect the limited set of object codes contained within the dataset. For example, the category "General Supplies" contains purchases such as workbooks. Although phonics curricula promote the use of workbooks as teaching materials, we cannot accurately determine specific purchases of items such as workbooks. Nor can we identify purchases of specific books (e.g., by ISBN) with these

[7]Figure A2 summarizes the years for which we obtained valid released third-grade exams and the number of questions by year.

data. A final limitation is that the data refer to expenditures at the school district–fiscal year level. We do not have information at the school or grade level. Thus, we cannot specify how Reading First funds were spent across specific schools or employees within a school district.

2.4 School District Tax Elections

To examine how the effects of Reading First expenditures compare to those of broader budget shocks on district spending and outcomes, we leverage data from Abott et al. (2020) on Texas school district tax elections to raise discretionary operational funds; these data cover 2000–2015, and we extend the record through 2019 using districts' adopted tax rates (see Appendix C.1 for additional details). Such elections generate plausibly exogenous increases in district revenues unrelated to any targeted programmatic objective. For this analysis, treated districts are those that passed a tax increase, and the comparison group consists of districts that had not yet passed one or never did. The analysis provides a plausible benchmark for us to assess whether the effects of Reading First reflect something specific about how those dollars were spent (e.g., on reading materials), not simply the infusion of additional resources alone. We merge the election data to PEIMS (discussed in Section 2.3) using district identifiers.

3 Empirical Approach

3.1 Difference-in-Differences

To examine the effects of Reading First on the extensive margin, we employ a difference-in-differences estimator with two-way fixed effects. We then use event studies to probe the underlying validity of our identifying assumptions. Rather than using actual school district participation in the program as a measure of treatment, we instead opt for an intent-to-treat measure. Our proxy for intent-to-treat is predicted program eligibility in the first year of the Reading First program.

Texas used a formula to determine eligibility for Reading First. The formula relied on five components, and for eligibility in a given year, the value of each component in the calendar year before a given funding cycle was considered (Texas Education Agency, 2003b). For clarity below, we consider just the first funding cycle, which is what we use as a proxy for eligibility for all districts.

Eligibility in the first funding cycle is broken into two groups. The first group contains only one component, the school district's third-grade reading exam failure rate on the 2002 standardized reading test being at or above the statewide average of 13.6%. The second group contains four components, and required that the school district had to oversee at least one of the following: (a) a geographic area designated as an empowerment zone or enterprise community under the federal tax code; (b) at least one campus identified for school improvement under Title I Part A; or a student population where (c) either 15% or more of students, or (d) more than 6,500 students, were eligible for basic Title I funding.

We summarize Reading First eligibility in cycle 1 for a district d as follows:

$$\begin{aligned}\text{Reading First Eligibility}_d = {} & \mathbb{1}(\text{Fail Rate}_{d,2002} \geq 13.6\%) \\ & \times \max\Big[\, \mathbb{1}(\text{Low Performing}_{d,2002}), \mathbb{1}(\text{Empowerment Zone}_{d,2002}), \\ & \qquad \mathbb{1}(\text{Share of Title 1 Students}_{d,2002} \geq 15\%), \\ & \qquad \mathbb{1}(\#\,\text{of Title 1 Enrolled Students}_{d,2002} \geq 6{,}500)\Big]\end{aligned} \tag{1}$$

where Reading First Eligibility$_d$ is an indicator variable equal to one if a district d is predicted to have been eligible for the program. We consider all districts ineligible in the years 2003 and earlier, and, conditional on being eligible, a district becomes eligible beginning in calendar year 2004. We use calendar year 2004 since this is the first post-funding receipt exam year.

Rather than compare districts that received money to those that did not, we compare districts that we predict to be eligible in the first year the program operated to districts we predict would be ineligible in this first year. We predict eligibility in this way to ensure that our treatment assignment is exogenous as all of the components that enter the program are from the calendar year before the program application had even been sent to the federal government. Some school districts that were ineligible in the first cycle, could become eligible for the second cycle as each subgrant cycle used the previous spring's third-grade reading fail rate relative to the state average in that year. Therefore schools that had not been funded in the first cycle used 2003 values to determine eligibility in the second cycle. The state-wide third grade reading exam fail rate in 2003 was 11%, which somewhat changed the set of schools eligible for funding (Texas Education Agency, 2003a, 2004).

We collected data on every component that is used in the formula of Equation 1. Despite this, our eligibility estimate is an imperfect proxy because there are slight deviations in the values in our collected data from the values the TEA was using at the time. We can perfectly reconstruct the same district-specific fail rate using our administrative data. But our estimates of low-performing school count, empowerment zone status, and Title I student information differ slightly from the TEA data that were being used.[8] Despite some minor differences in the values for the non-fail-rate components, our predicted eligibility is quite a good measure of actual eligibility and funding receipt. For the first cycle, 90% of the districts who received Reading First funds were deemed eligible by our measure, and 55% of those districts that were funded in the second cycle were likewise considered eligible using our 2002 criteria. Therefore, few districts we classify as ineligible using the 2002 criteria ever receive Reading First funds (5 percent) and 75% of these became eligible under the cycle 2 criteria, which used the 2003 test threshold rather than the 2002 exam threshold.

While we think it is unlikely that school districts were endogenously manipulating their own fail rate in 2003, a month before the state application was submitted to the federal government, we find it more palatable from an identification standpoint to restrict all characteristics to their

[8]The TEA published a list of eligible districts, along with values for some of these components for each eligible district. They did not publish the values of each component for ineligible districts.

2002 baseline level. This decision will also attenuate our treatment effect estimates, biasing us away from finding positive treatment effects. This occurs because some districts we classify as ineligible actually received treatment while other districts that we deem eligible never ended up receiving program funds.

Our main difference-in-differences estimating equation is as follows:

$$\begin{aligned} y_{cg(d)t} = \alpha_{cg} + \gamma_t + \beta\, \text{Reading First Eligibility}_d \times \mathbb{1}\{\text{Year} \geq 2004\} \\ + \sum_{\substack{t=1995 \\ t \neq 2002}}^{2017} \psi_t \left(X_c^{2002} \times \mathbb{1}\{\text{Year} = t\} \right) + v_{cg(d)t} \end{aligned} \tag{2}$$

Here $y_{cg(d)t}$ is our outcome of interest. For our initial analysis this is the mean standardized reading test score in grade g on campus c of school district d in year t. The terms α_{cg} and γ_t represent campus–grade and year fixed effects.[9] Reading First Eligibility$_d$ is a binary variable interacted with the post-Reading First treatment, indicated by $\mathbb{1}(Year \geq 2004)$. Thus β is our treatment effect estimate of interest.

For long-run outcomes, we move from estimating campus–year effects to following exposed cohorts through high school, college, and the labor market (Section 6). We adapt the specifications above to the student–cohort level, defining treatment using the district and year in which each student attended the third grade. For test score, high school, and college outcomes, we estimate fixed effects models analogous to Equation (2). For labor market outcomes, we estimate fixed effects Poisson models, which model zeros explicitly (Chen and Roth, 2024).

Our main specification also controls for baseline campus composition by interacting the campus share of economically disadvantaged students in 2002, X_c^{2002}, with year fixed effects. We do this because Reading First targeted poor districts and these interactions allow test scores to trend differently by baseline poverty. This is included mostly for identification purposes, since there may have been differential trends by underlying poverty or different programs or incentives that differentially affected poorer districts occurring across our study period. In the appendix, we show that our results are similar, if not larger, if we exclude this control. We also find similar estimates if we include additional baseline campus demographic controls or control for time-varying campus demographics. We weight each regression by the number of students who take the grade g exam in year t at campus c, so that the treatment effect estimate represents the effect for the average treated student. For all specifications, confidence intervals are constructed from robust standard errors clustered at the school district level.

We estimate simple modifications to this basic specification in various places throughout the paper. The most important is the one that restricts the set of school districts included in the analysis to be those who had a 2002 third-grade reading score fail rate closer to the state average fail rate. We do this because it is possible that districts that had a very high failure rate on the 2002 third-grade reading exam are not comparable with those school districts

[9]For analyses where we restrict to a particular grade (e.g., the third grade), the g subscript in the fixed effects becomes redundant.

that had a very low failure rate. We refer to this restriction as the "bandwidth" specification and use 5 percentage points as the bandwidth. We explore how our treatment effect varies with a variety of different bandwidths in the appendix. Given that Reading First funding ended in 2011, we also estimate a version of Equation 2 that allows the treatment effect to differ across the two periods, replacing $\beta\text{Reading First Eligibility}_d \times \mathbb{1}(Year \geq 2004)$ with $\beta^d\text{Reading First Eligibility}_d \times \mathbb{1}\{2004 \leq t \leq 2011\} + \beta^a\text{Reading First Eligibility}_d \times \mathbb{1}\{t > 2011\}$, where β^d captures the effect during the program years and β^a the effect in the years following its conclusion.

We amend these equations to fit other outcomes accordingly. For instance, when we examine 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, expenditures, or employment we estimate our regressions at the district–year level.

Our identification strategy relies on the assumption that test scores in eligible districts would have evolved in a similar manner relative to test scores in untreated districts had the treatment not occurred. Since we cannot observe the counterfactual outcomes of treated districts in the absence of the program, we approximate a test for this assumption by exploring the relative trends between treated and control units before the program began. In our setting, campus fixed effects, account for all time-invariant cross-sectional differences between schools and districts, and year fixed effects account for things that affect all campuses and districts across time. The inclusion of the poverty-specific time trends also allows poorer schools to be on differential trends or have differential shocks relative to other schools.

We investigate this assumption in two ways. First, we compare raw differences in mean reading scores in eligible districts to ineligible districts relative to what their difference was in 2003. Second, we employ an event study approach using the following equation:

$$\begin{aligned} y_{cg(d)t} = \alpha_{cg} + \gamma_t + \sum_{\substack{t=1995 \\ t \neq 2003}}^{2017} \beta_t \,\text{Reading First Eligibility}_d \times \mathbb{1}\{\text{Year} = t\} \\ + \sum_{\substack{t=1995 \\ t \neq 2002}}^{2017} \psi_t \left(X_c^{2002} \times \mathbb{1}\{\text{Year} = t\} \right) + v_{cg(d)t}, \end{aligned} \tag{3}$$

The coefficients of interest in Equation 3 are β_t, capturing the relationship between Reading First eligibility and test scores over time, both before and after a Reading First subgrant. To explore the identification assumption, we will consider if there are differences in pre-policy trends, captured by β_t for $t \leq 2003$. Similarly, the coefficients β_t for $t \geq 2004$ will allow us to examine the dynamic effects for the years after Reading First was introduced. Of note, event time t runs from calendar year 1995 to calendar year 2017, with $t = 2003$ as the omitted reference period. Since we consider every eligible district as treated in the same year, event time and calendar time are identical.

3.2 Instrumental variables approach

Conditional on being eligible for funding, Texas used a separate formula to determine the minimum amount of money each school district should be awarded from Reading First if funded (Texas Education Agency, 2003b). That formula forms the basis of our simulated instrumental variables approach as we use it to predict the amount of money each school district spent on the program.

Since larger districts might require more money to operate the program than smaller districts, Texas decided that each funded district should receive at least the same proportion of the state's total Reading First funds as the share of the state's Title I Part A funds the district received 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.

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

$$\begin{aligned}\text{Reading First Revenue}_{dt} &= \text{Reading First Eligibility}_{d} \\ &\times \max\left(\$170{,}000,\ \text{Reading First State Funds}_{t} \times \frac{\text{Title 1 Funds}_{d,2002}}{\sum_{j=1}^{N} \text{Title 1 Funds}_{j,2002}}\right) \qquad (4)\end{aligned}$$

Here, eligibility is determined by Equation 1; Reading First State Funds$_t$ is the portion of Texas's federal Reading First appropriation in year t that was allocated to districts (80% of the total); N is the number of Texas districts with a Title I Part A allocation in the base year (2002); and Title 1 Funds$_{d,2002}$ is the amount of Title I funds that a district d received in the base year (2002). As with our eligibility proxy in Equation 1, we opt to hold everything fixed to values in the base year 2002. Within this prediction, the only component that varies across time is the Reading First funds awarded to the state of Texas by the federal government and the only component that varies across school districts is their share of title I funds allocation in year 2002.

Our preferred measure of Reading First exposure for a given district d in year t is the cumulative Reading First expenditure that the district has spent up until time t.[10] Cumulative predicted revenue, constructed in Equation 4, is the instrument for our endogenous variable, cumulative Reading First expenditure. We prefer cumulative expenditure for two reasons. First, this is the most comparable with the difference-in-differences and event-study estimates, which consider a post-treatment period. If the effects are due to changing pedagogy or the use of durable assets like books, then any estimated treatment effects should not immediately

[10] An alternative to cumulative expenditure is cumulative expenditure per student, or even a simple binary instrument for ever-eligible and ever-spent. We find similar results if we either scale by the number of K-3 children enrolled in a district or collapse the instrument to a simple binary measure. In either case, the sign and significance pattern of the instrumental variables estimates are essentially unchanged.

collapse to zero the moment funding ends. Second, districts did not have to spend every received Reading First dollar in the same calendar year the federal government sent that money to the state. As shown in Figure 1B districts often received federal money in one fiscal year and spent it in the following years.

The following equation provides the first stage of the instrumental variables specification:

$$\sum_{s=2004}^{t} \text{Reading First Expenditures}_{ds} = \lambda_{cg} + \delta_t + \pi \sum_{s=2004}^{t} \text{Reading First Revenue}_{ds} + \sum_{\substack{t=1995 \\ t\neq 2002}}^{2017} \phi_t \left(X_c^{2002} \times \mathbb{1}\{\text{Year} = t\} \right) + \varepsilon_{cg(d)t} \tag{5}$$

Similar to our difference-in-differences specification, we include campus–grade and year fixed effects, interacted controls for school poverty levels, and cluster our errors at the district level. Both expenditures and predicted revenue, and their sums, are equal to zero before the year 2004. The coefficient π measures how many Reading First dollars a district actually spent for each received dollar the formula predicts.

In the second stage, we replace actual expenditures with their fitted values from Equation 5 and estimate:

$$y_{cg(d)t} = \alpha_{cg} + \gamma_t + \theta \left(\sum_{s=2004}^{t} \widehat{\text{Reading First Expenditures}}_{ds} \right) + \sum_{\substack{t=1995 \\ t\neq 2002}}^{2017} \psi_t \left(X_c^{2002} \times \mathbb{1}\{\text{Year} = t\} \right) + v_{cg(d)t} \tag{6}$$

where $\sum_{s=2004}^{t} \widehat{\text{Reading First Expenditures}}_{ds}$ captures the fitted cumulative Reading First expenditures from the first stage in Equation 5. 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 \$20 million (2025 dollars), roughly the cumulative Reading First dollars received by the average eligible student's district over the program (\$18 million). Similar to our difference-in-differences estimate, we weight each observation by the number of students who took exam grade g in year t in campus c. Standard errors are adjusted to account for the fact that this is an instrumental variables specification. All other variables are as specified in Section 3.1.

Our instrument is a simulated variable constructed from the eligibility formula, in the simulated-instrument tradition of Currie and Gruber (1996). Our prediction considers each district's 2002 characteristics and then applies the programmatic rules to those fixed characteristics, which is similar to the formula instruments used to study Title I and school-finance

reforms (Gordon, 2004; Cascio et al., 2010; Jackson et al., 2016; Lafortune et al., 2018).[11] Identification therefore comes from the program's rules, the 2002 fail-rate cutoff, the poverty criteria, and Title I shares frozen in 2002, rather than from comparison of poorer districts with richer ones, or larger and smaller districts. Thus, no action a district took, such as changes in enrollment, spending, or student outcomes, enters the instrument.

Conditional on the fixed effects and controls, the exclusion restriction is that predicted Reading First revenue affects outcomes only through actual Reading First expenditure. Level differences in size and poverty are absorbed by the campus fixed effects, and differential trends by poverty are absorbed by the 2002 economically disadvantaged share interacted with year fixed effects. The latter interaction lets every school with a different poverty level have its own specific time path, which can flexibly soak up differential trends. This ensures that variation in our instrument comes from changes in the predicted revenue and not from the level pre-existing need in each school.

This interaction, the narrow-bandwidth, and the heterogeneity analyses in Section 4.3 eliminate many potential threats to identification. These heterogeneity analyses restrict the set of either schools or students in both the control and the treated groups to have the same characteristic (e.g., above median Title 1 share). Thus a finding of similar effect size within comparisons of the most similar students or schools helps dissuade any concerns that our results simply reflect that poor schools might have been differentially eligible and on differential trends. We also explore if predicted Reading First revenue is predictive of other forms of funding after 2003, and find no evidence that Reading First predicts non–Reading First revenue (Appendix Table B4) and that our main estimates are robust to controlling for non–Reading First district revenue sources (Appendix Table B5).

Finally it is worth restating that Reading First revenues and expenditures are only observed at the district–year level. Thus, in our specifications, every campus in a district is assigned the same revenue and expenditure. A district that allocated its Reading First revenue on a subset of campuses is treated identically to one that spread funds evenly. Again, we do examine whether higher-poverty campuses, or students with different characteristics within a district gained more than their untreated counterparts.

Our two main estimators complement each other but likely measure different aspects of Reading First. The difference-in-differences captures the extensive-margin change—the effect for the average treated student of a district's becoming eligible. The instrumental variables estimate captures the combination of this extensive margin effect and the intensive margin—that of an additional dollar. If eligibility had mattered only through the dollars it brought and every dollar had the same effect, the two sets of estimates would perfectly coincide. However, we have reasons to expect the effects at each of these margins to differ and even to expect nonlinear effects due to additional dollars. Part of the treatment arrived with binary changes rather than things that scale with extra money. For example, Reading First induced curricula

[11] Our instrument is also related to instruments that apply a changing policy schedule to a baseline input fixed in an earlier year (Gruber and Saez, 2002; Dahl and Lochner, 2012).

changes may be independent of the size of the grant a school district received. Moreover, there may be diminishing returns to an additional dollar spent. Last, the two estimators weight districts differently since the variation in dollars is concentrated among the largest grants. We return to this distinction when presenting our main findings.

3.3 Other threats to identification

Our design requires that eligibility not coincide with other changes affecting reading outcomes. The clearest candidate threats among such changes are the rest of No Child Left Behind or changes in Title 1 funding. If the non-Reading First provisions of No Child Left Behind or Title 1 had their own separate effect on improving reading instruction and these changes were correlated with Reading First treatment, our estimates would not clearly identify the effect of Reading First. Our analysis limits this concern in a few ways. First, many of these provisions applied statewide. All Title I schools, in eligible and ineligible districts alike, faced sanctions based on their own performance, and Reading First eligibility did not itself trigger any sanction. Second, our narrow-bandwidth specifications compare districts near the 13.6% state-wide 2002 fail rate, who are plausibly more similar to one another than districts farther away from the cutoff. Third, our heterogeneity analysis compares only similar schools and students to one another, including subsets such as just examining the poorest schools and students. Finally, we consider a placebo test for the effect of eligibility on revenue outside of Reading First (Appendix Table B4) and estimate specifications that control for other revenue and for the 2002 socioeconomic composition of the campus interacted with year fixed effects.

Another potential confounder was the introduction of the Texas Student Success Initiative in 2002. Beginning with the 2002–03 third-grade cohort, this initiative required students to pass the state reading test within three attempts to advance to fourth grade. Zhong (2026) shows that retention under this policy temporarily raised retained students' subsequent scores while reducing high school graduation and adult earnings. The Student Success Initiative was a student-level policy, not a campus- or district-level policy and affected all third graders in the state of Texas equally. However, given that the initiative was introduced one year before our Reading First outcome data begin, we explore this potential threat to validity, as it remains possible that reading-first-treated districts had larger shares of students who were affected by the initiative. We verify that the initiative does not drive our estimates by exploring initiative-specific money, targeting, and retention. Further because this initiative mandated retakes for students who failed, we only consider each student's first attempt in our analysis. Finally, we find similar effects when including all tests taken in a given year in Appendix Table B6.

3.4 Predicted Eligibility and District Revenues

In this section, we perform a mechanical check of whether our predicted eligibility and predicted revenue are related to actual Reading First revenue. We begin by establishing that Reading First eligibility did indeed increase Reading First related revenue. Table 1 presents

the relationship between Reading First eligibility and Reading First–specific revenue, which is trackable using a fund code specific to Reading First in the annual district level finance data. Thus the unit of observation here is a school district-year. Panel A presents difference-in-differences estimates for the entire post-period. Panel B splits the post–Reading First period into separate joint treatment effect periods for the years during the Reading First program (2004–2011) and the post-2011 years, once the funding had ended. While, Panel C uses our predicted Reading First revenue from Equation 4 to estimate actual Reading First revenue. Column (1) in Table 1 considers all school districts, and Column (2) restricts the analysis to school districts within five percentage points of the 13.6% reading test score fail rate (in 2002) used in the eligibility formula in Equation 1.

Across all specifications, we see that Reading First eligibility is predictive of a district receiving Reading First revenue. Column (1) Panel A shows that relative to ineligible districts, Reading First–eligible districts received an additional $75 per K–3 student post-2003. Column (2) Panel A suggests that the estimated effect is similar ($59 per K–3 student) when we restrict the analysis to districts within five percentage points of the 13.6% fail rate. In Panel B, we see that the estimated increases in Reading First revenues are confined to the period when program funding was operational, 2004–2011.[12] Over this time period, Reading First–eligible districts received $132 per K–3 student. Following the program's end, the revenue effect drops to a precise null effect. Finally, Column (1) Panel C shows a strong relationship between predicted and actual Reading First revenue. Each additional $1 million in cumulative predicted Reading First revenue is associated with a $1.41 million increase in actual cumulative Reading First revenue.

4 Test Scores

4.1 Third Grade

We begin by presenting raw data demonstrating the relative difference in standardized third-grade test scores across time. Figure 2 Panel A shows the difference between eligible school districts and ineligible school districts across time, relative to their difference in 2003. This figure has two descriptive patterns of interest. First, prior to the program start in 2004, the difference between eligible and ineligible districts was stable across time and had no discernible trend. Second, nearly immediately following the beginning of the program, reading scores rose in eligible districts relative to ineligible districts. These effects appeared to grow across time and then level out beginning in 2011.

Having discussed the descriptive patterns of interest, we now turn to present difference-in-differences results from our main estimating Equation 2, which are presented in Table 2. Column (1) of Panel A reports the pooled difference-in-differences estimate, indicating that

[12] We stop here in 2011 because this is two years after the federal government's final payment to the state of Texas. Typically, the state received the money in the year after the federal money was appropriated, and districts sometimes took an additional year to spend the money.

school performance on third-grade reading exams in eligible districts improved by 0.11 standard deviations after the program began. Panel B separates the treatment effect into the active funding and post-funding periods. The estimated effect is 0.07 standard deviations for 2004–2011 and 0.16 standard deviations for the years after 2011. Across all of these panels, we find similar effects for those schools within the bandwidth. This pattern is consistent with the fact that an increasing share of each cohort was treated across time as illustrated in Panel D of Figure 1. It was not until 2008 that all treated students were fully treated.

We next examine the event study estimates from the Specification 3, which are presented in Figure 2 Panel B. The pattern of results in the figure confirms the absence of pretrend differences between Reading First–eligible and ineligible districts. We see a clear trend break after the start of Reading First in 2004 and a positive effect on reading test scores. The estimated effect increased until 2012, one year after the last Reading First dollars were spent. For the years after this, the effect levels out and remains positive and statistically significant. Figure 2C reestimates our main event study specification but restricts the analysis to school districts within five percentage points of the 2002 13.6% fail rate used to determine eligibility for Reading First. This specification allows us to compare districts plausibly more similar at baseline. The bandwidth model shows a similar pattern.

Panel C of Table 2 displays the results from the instrumental variable specification, where predicted revenue is an instrument for actual expenditures. Column (1) reports the estimate for the full sample. Here the first-stage F-statistic far exceeds the conventional weak-instrument threshold (F = 1,711), and Appendix Table B3 shows that the first stage is similarly strong across different robustness specifications (F-statistic between roughly 1,200 and 2,700 in the full sample and above 750 in the narrow-bandwidth sample). As with our difference-in-differences model, we see that additional cumulative Reading First revenues improved third-grade reading. The average treated student at a treated school district, was in third grade at a time when that school district received, on average, $18 million in Reading First funding (in 2025 $). Thus, for comparability with the difference-in-differences estimates, we scale our reported instrumental variable estimates by $20 million.

Our instrumental variables estimates are smaller than our difference-in-differences estimates. At the average eligible district's cumulative funding ($20 million), the instrumental variables reading effect is about half as large. The gap reflects what each estimate measures. The difference-in-differences captures the effect of eligibility for the average treated student, and the instrumental variables estimate captures the effect of the average dollar. Appendix Figure A4 estimates both models separately for eligible districts in three groups of cumulative predicted funding (under $5 million, $5 million to $15 million, and $15 million or more). Districts in the high-spending group received eighteen times the funding of those in the low-spending group. Although the reading gains were higher for the high-spending group, they were only about two and a half times larger, not eighteen times. Panel A shows the treatment effect by group at each group's average funding level for a treated student within that group. Panel B rescales each estimate to be at the same dollar level. On a per dollar basis, the reading

gains decline monotonically from the low-spending to the high-spending group. Within each spending group, however, the difference-in-differences and instrumental variables estimates are very close to one another. The pooled estimates differ because the local average treatment effect is dominated by the high-spending group. This group contains the most students and the highest average spending. Here, each additional dollar produced the smallest gains. The low-spending group's per dollar estimates are more imprecise and have a weaker first stage as a larger share of the eligible districts that were never actually awarded Reading First dollars are contained within this group.

We also consider whether Reading First, beyond raising average scores, shifted students across the achievement distribution. Figure A5 plots the effect of Reading First eligibility on the probability that a first-time third-grade test taker places in each decile of the statewide reading distribution. Eligibility reduces membership in the bottom deciles—the student's probability of placing in the lowest decile falls by roughly three percentage points—and increases membership in deciles seven through nine by one to three percentage points. The pattern is similar when we restrict to districts within five percentage points of the eligibility threshold and when we consider shifts in the reading decile distribution of eighth graders. True to the program's aim of targeting struggling readers, this evidence indicates that Reading First moved struggling readers out of the bottom of the distribution rather than merely boosting students who were already proficient. This finding is consistent with work showing that reading programs produce their largest gains among students most likely to struggle (Machin et al., 2018; Banerjee et al., 2007; Andersen et al., 2026). Figure A5 also examines the probability of a student transitioning across reading test score deciles between third and eighth grade. For example, the very top left-hand corner cell suggests that conditional on being in the bottom decile of a reading test score, a Reading First third grader was 1.1 percentage points less likely to remain in that lowest decile by eighth grade than a non–Reading First student. Overall, Panel C of Figure A5 provides evidence suggesting that low-performing third-grade students had increased probabilities of moving to higher test score deciles by eighth grade.

We also examine the effect of Reading First on math in Figure A6 and find improvements in math test scores. In standardized terms, according to our main equation (Equation 2), the math effect is 0.11 standard deviations, almost identical to the reading effect (see Appendix Table B7). 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 5.1, we investigate the potential mechanism that leads to this result. Specifically, we assess the plausible connection between improvement in reading and math competency by examining whether student success differs by the number of words in standardized math test questions. Appendix Figure A7 repeats the decile analysis for math and shows a similar pattern, with students moving out of the bottom decile and into the upper-middle deciles.

Finally, we examine behavioral outcomes that might be influenced by improved reading performance: disciplinary actions, suspensions, and attendance. As indicated by the event studies in the panels of Figure A8, we do not find an effect of Reading First on those outcomes.

4.2 Internal Validity Checks

Panels D and E of Figure 2 present internal validity checks for our difference-in-differences and instrumental variables models, respectively. Our results are quite robust across this battery of specification checks. Every row of each panel corresponds to a separate specification, with the first row representing our baseline estimate. Row 2 removes the economically disadvantaged control from the baseline specification, leaving no controls in the regression. Row 3 adds 2002 demographic controls (percentage Black, percentage Hispanic, percentage Asian/Native American) interacted with year fixed effects. Row 4 replaces those interacted controls with time-varying demographics. Row 5 uses the same demographic controls as row 3 and adds the share of students who are gifted, or ESL, or who receive special education. Row 6 removes the balanced campus restriction and includes all campuses, which results in an unbalanced panel of schools. Row 7 removes any student weights. Texas school districts lie within 20 Education Service Center regional units, which provide services to improve school performance.[13] Given that Reading First districts interact with these regional centers and that this interaction may foster spillovers to other districts, row 8 clusters standard errors at the region instead of district level, and row 9 includes region–year fixed effects. During our sample period, Texas changed how exams were structured two times.[14] Row 10 excludes these exam transition years. Row 11 keeps years before 2012, when Reading First funding was available. The last row includes all test takers, not just students taking the reading exam for the first time. Appendix Figure A9 plots the full event study for each of these specifications and reports positive effects on third-grade reading similar to those from the baseline specification.

Figure A10 explores the sensitivity of our bandwidth subset analyses to alternative choices beyond 5%. Overall, our results are positive irrespective of bandwidth. The effects increase with the bandwidth up to about twelve percentage points and then level off, with the difference-in-differences estimates less precise for the narrowest bandwidths of three percentage points and below. The first-stage relationship is also stable across every robustness specification, with instrument F-statistics well above conventional weak-instrument thresholds (Appendix Table B3).

We also contend with the possibility that districts received other revenue at the same time that Reading First funds arrived. Reading First was one program within the broader No Child Left Behind Act, and the district characteristics that determined eligibility may have also shaped exposure to other government funding streams. If that was indeed the case then our estimates of the treatment effect of Reading First would be affected by those other funding sources. To address this concern we first test whether Reading First led to changes in non–Reading First revenues. Although Reading First may have correlated with district factors

[13] 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 and serve as intermediaries between districts and the TEA.

[14] 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.

or have led to changes that could have subsequently affected a school district's revenue, a finding of little or no effect on non–Reading First revenue would lend additional credibility to our identification strategy and the argument that improved reading instruction is the channel driving our findings. We present the effects of Reading First on non–Reading First revenue in Appendix Table B4. We find no economically or statistically meaningful effect of Reading First on non–Reading First district revenue. Second, we reestimate our main specification while controlling directly for non–Reading First revenues and find that our estimates are unaffected (Appendix Table B5).

We also address the possible threat of selection-through-retention. Starting in 2002–03, Texas introduced the Student Success Initiative, which retained third graders who did not pass the state reading exam. Zhong (2026) studies this policy directly and finds that retention temporarily raises the test scores of retained students. Therefore, average scores might rise mechanically as the mix of test takers changes, particularly if there are relatively more repeat test takers in Reading First districts. We address this issue by restricting the sample to first-time test takers, which removes any retake-driven change in composition. We also verify that Reading First itself did not change retention behavior (see Figure A11). As a final check, we reestimate our main results on the sample of all test takers in Table B6 and Figure A12 and find effects very similar to those from our baseline specification. We also explore whether Reading First districts were differentially exposed to Student Success Initiative-related funding, and find no such evidence.[15] Reading First eligibility does not have any predictive power on whether a district received any funds from this initiative nor how much it received. The dollar estimate for those funds is statistically insignificant and negative near the 13.6% fail rate threshold, and the event studies show no differential movement in Student Success Initiative funding as a result of Reading First's rollout (see Appendix Table B2 and Appendix Figure A1).

4.3 Heterogeneity Analysis

Figure 3 summarizes heterogeneity in the Reading First effect along two dimensions. Panels A and B examine variation across campuses based on 2002 campus characteristics, and Panels C and D examine variation across students within campuses based on individual characteristics. Within each specification, we restrict the analysis only to schools or students who meet the reported criteria in the analysis, dropping all other observations. For each dimension, we report both difference-in-differences estimates and the corresponding instrumental variables estimates per $20 million of Reading First dollars (2025 dollars) for the full sample and for districts within five percentage points of the eligibility threshold.

Beyond documenting heterogeneity in treatment effects, the exercise in Panels A and B of Figure 3 has the added benefit of comparing campuses within Reading First and non–Reading First districts that are more similar to one another along observable dimensions. Given that a

[15] The Student Success Initiative provided state grants for accelerated instruction, such as tutoring and summer school, for students at risk of failing the promotion exams (Curry, 2005).

central goal of Reading First was to improve reading outcomes in high-poverty schools, we are particularly interested in whether the treatment effects are larger for campuses serving more disadvantaged populations. We begin with the difference-in-differences estimates in Panel A. Rows two through five examine heterogeneity by racial and ethnic composition. Campuses with above- and below-median shares of Black students show similar point estimates, as do campuses split by Asian/Native American enrollment share. The clearest difference by race or ethnicity appears for the Hispanic share,[16] where campuses above the median show a notably larger effect than those below it. The next row corroborates this pattern from the other direction, as campuses with above-median White enrollment show a smaller effect.

Rows six and seven turn to measures of economic disadvantage. The Reading First effect is larger for campuses with above-median Title I shares and above-median shares of economically disadvantaged students, while the corresponding below-median estimates are smaller. Thus, at the extensive margin, we find that Reading First had its largest effects in higher-poverty schools with larger shares of disadvantaged students. These are the students the funding formula most directly targeted, so the program appears to have reached the population it was designed to serve. In Panel A, we find similar results when stratifying by the share of students classified as special education, gifted, or ESL. The intensive-margin estimates in Panel B are more uniform across estimators. The per-dollar improvements in reading scores are similar for campuses above and below the median Title I share. Thus, we see the concentration of effects among higher-poverty campuses at the extensive margin rather than for the marginal dollar.

Panels C and D stratify by student characteristics. The point estimates are positive for every racial and ethnic group. In the difference-in-differences specification, the gains are broadly similar across racial groups. In the instrumental variables specification, the per dollar effects are largest for Hispanic students and relatively smaller for Black and White students. We also consider specifications whose outcomes are the difference between Black and White or Hispanic and White test scores. Consistent with these patterns, the gap outcomes at the bottom of Panels C and D show a narrowing White–Black gap in the difference-in-differences specification and a narrowing White–Hispanic gap in the instrumental variables specification, which provides suggestive evidence that Reading First reached the disadvantaged populations it targeted. The positive effects of Reading First on reading are similar for both male and female students, while the estimates for ESL students are imprecise. Consistent with the campus-level heterogeneity analysis, economically disadvantaged and Title I students show relatively larger test-score gains. Special education students also show large gains, in line with the program's objective of reaching struggling readers, with gifted students gaining less.

Last, motivated by a growing literature documenting the effects of air pollution on both short-run and long-run student performance, Table B8 splits campuses by median $PM_{2.5}$ concentration (Currie et al., 2009; Ebenstein et al., 2016; Hollingsworth et al., 2022; Persico and Venator, 2021; Jacqz, 2022; Yoder, 2024). The difference-in-differences estimates are similar across the two groups. The instrumental variables estimate is larger for campuses with below-

[16] Hispanic residents were the largest minority in Texas during our sample period.

median pollution, but it rests on a much weaker first stage. We note that pollution exposure is correlated with economic disadvantage and demographic composition in this setting.

5 Mechanisms

5.1 Decoding Test Questions

As described in Section 2.2, we digitized third-grade reading and math exams and classified each question by its word count and by the share of its words that can be decoded with basic phonics skills. For reading exams, we also determine whether each question tests vocabulary or broad comprehension. For math exams, we create an indicator for answers that only include numeric choices. There are almost no math questions that are entirely free of words in the question content, so this serves as an approximation for whether or not the question is pure computation or more of a word problem.

We examine the effects of Reading First on the share of correct responses on third-grade exams at the question–campus–year level as the outcome. Table 3 presents the results. Columns (1)–(2) pool the math and reading exams, columns (3)–(4) report the reading exams, and columns (5)–(6) report the math exams, in each case for all districts and for the narrow-bandwidth sample. Panel A of Table 3 reproduces our main finding using the share of questions answered correctly rather than standardized scores. We take the average share of correct answers in every question of a third-grade reading or math exam by campus–year, which makes our unit of observation the campus–grade–year–exam question. Consistent with our earlier results, the estimates show that Reading First students outperformed non–Reading First students by 2.3 percentage points in reading and by 2.3 percentage points in math.

Each row of Panel B comes from a separate regression that adds one question feature and reports the interaction of Reading First eligibility for the entire post-treatment time period, and an indicator for that particular question feature. The estimated regressions also include all other interactions and indicators between Reading First eligibility, post-treatment period, and feature presence, which we do not report for ease of interpretation. The first row interacts the treatment with the word count of each question. Reading First eligible students score 0.04 percentage points better per word in questions with more words relative to their non-Reading First peers. A question that is one standard deviation longer (about 25 words) carried an additional advantage of roughly one percentage point, comparable to the average effect itself. Across all samples treated students do better on questions where phonics skills are more applicable. Specifically, Reading First students do around 0.02 to 0.03 percentage points better on a question with one percentage point more words that can be decoded with basic phonics skills (CIERA levels 1–4; see Section 2.2). The treatment effect estimate is consistent across the pooled, reading, and math questions. A one-standard-deviation (ten-point) difference in the decodable share is about ten percentage points, which would translate to an improvement of about 0.3 percentage points on the average question.

For reading questions only, columns (3) and (4) show that this advantage did not extend to vocabulary questions, which test whether students know the meaning of specific words in a reading passage. However, Reading First students did much better on questions requiring 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. These gains are what we would expect if better decoding skills free up attention for understanding the passage itself (Gough and Tunmer, 1986; Andersen et al., 2026).

Overall, the reading results suggest that Reading First strengthened a foundational skill rather than performance on a particular exam. If this is the case, we would expect the gains to appear on other tasks requiring students to read. The math exam results in columns (5)–(6) of Table 3 allow us to examine this hypothesis. Consistent with our earlier math test score results, Reading First eligibility increased the share of math questions answered correctly by 2.3 percentage points, an effect comparable to that on reading. The composition of the two exams may explain this comparability of effect sizes across domains. Many math questions are word problems, and a student who cannot parse the question will do worse on these assessments (Abedi and Lord, 2001; Fuchs et al., 2015). Reading skill is therefore an input on much of the math exam, so an improvement in reading should raise math scores, as well.

Panel B indicates that reading is a channel behind the math gains, though not the only one. As in reading, the advantage grows with the length of the question (0.04 percentage points per word) and with the share of words that require basic phonics skills. We also explore whether students treated by Reading First did better on the most computationally intensive questions. Almost no math question has zero words, but the answer choices available to students often have zero words. For roughly 28 percent of questions, each of the four answer choices is entirely numeric. We use these questions as a proxy for the most computationally intensive items, as opposed to those that require reading and comprehending a word problem. This is, of course, an imperfect proxy, but it is straightforward to estimate given the digitized exams and, since it is a common question type, it has reasonable support in the data. Reading First treated students did not do any better on these computationally intensive questions compared to their untreated peers.

In general, the math gains are concentrated in the question types where more reading is required, and not concentrated in those where more computation is required. This pattern is consistent with recent work that documents the spillovers of targeted interventions across different subjects (Xu et al., 2023). It is also important to note that this pattern of results is difficult to reconcile with an alternative explanation that math scores simply increased because eligible districts received more resources. If this were the case, we would expect improvement on computationally intensive math questions. What we find instead is that improvements were achieved only on math questions that relate to reading and decoding ability.

5.2 Expenditures

Since we find positive effects of Reading First on student test scores, our next step is to explore the compositional allocation of the funds. Figure 4 summarizes the allocation of Reading First expenditures using our main difference-in-differences model from Equation 2. Each point is the estimated effect of eligibility on Reading First spending in the indicated category per K–3 student during the program years, 2004–2011. The four panels form a hierarchy. The first reports total tracked expenditures, and each subsequent panel decomposes the last category of the panel before it, so each panel has its own dollar scale. Panel A shows that eligible districts spent $131.47 more per K–3 student from their Reading First funds relative to ineligible districts. We find similar, if somewhat smaller, effects when we restrict the comparison to districts within five percentage points of the 2002 fail rate cutoff of 13.6%.[17]

Panel B breaks up the total expenditure reported in Panel A into its broad components. The largest shares went to salary, wages, and benefits ($72.50) and to supplies and materials ($46.78). Districts also spent modestly more on professional purchased services ($8.35) and travel ($2.68).[18] Given that supplies and materials are the second largest category, Panel C decomposes this spending into its individual components: general supplies, testing materials, and books. The increase is split roughly evenly between books ($23.90) and general supplies ($22.43), with a small and statistically significant effect on testing materials ($0.41). Panel D further disaggregates the books category into reading materials, library books, and textbooks. Reading materials account for nearly all of the book spending result, $22.25 of the $23.90 total. The effects on library books ($0.31) and textbooks ($1.34) are statistically significant but economically small. The limited response of library books and textbooks may not be surprising given the program's focus on teaching students how to read, which presumably includes existing books. Library books and textbooks would not have become obsolete simply because a district joined Reading First. Therefore, these results are consistent with spending on phonics-based materials responding through the reading materials category, which is likely where phonics-related curricula would be cataloged.

Another possibility is that the increase in reading materials expenditures was used to update existing reading materials, independent of phonics instruction. Thus, the reading improvements we find may be due to updated or new materials, not curricular content. However, Appendix Table B9 shows that these expenditures were not sustained beyond 2011, when Reading First funding ended, yet the third-grade test score effects persisted, consistent with lasting changes in instructional practice. Moreover, recent evidence on math textbooks suggests that new materials on their own are unlikely to explain these gains. Across six states that adopted the Common Core standards, schools using different elementary math textbooks saw similar achievement gains, and Blazar et al. (2020) conclude that adoption of a new textbook alone

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

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

is unlikely to improve student outcomes. Reading First instead changed the type of reading curriculum and paired the new materials with coaching and teacher training. The durable nature of books indicates that schools may have also purchased phonics-related materials and used them well beyond the end of the Reading First program in 2011 or that funds that would have otherwise been used to purchase non-phonics-based materials were now allocated to resources backed by the science of reading.

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

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

The salary estimates also preview the comparison we explore in Section 5.4, where we show that discretionary revenue from local tax elections flows predominantly to school payrolls, with no detectable increase in spending on reading materials, and produces relatively smaller reading gains per dollar than Reading First. The difference between unconditional funding and a more pedagogically targeted approach is suggestive evidence that the Reading First spending on payroll was effective beyond the wage and retention channels documented in the teacher pay literature. Better teacher pay can improve student outcomes by helping districts recruit and retain teachers (Loeb and Page, 2000; Hendricks, 2014; Biasi, 2021). But even payroll-heavy Reading First awards came with the program's curricular mandates, coaching, and professional development, which may have made these dollars effective beyond those channels.

5.3 Teacher and School Characteristics

Another mechanism that could explain the effect of Reading First would be changes in the composition of campus teachers. For instance, if higher-quality teachers moved to Reading First districts post-2004, then the effects we find might be attributable to teacher-specific quality instead of 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 do not allow us to link students to teachers through classroom identifiers for years before 2012 (after the Reading First funds ended). As an alternative, we follow Foy (2024) and estimate campus value added using student-level data from 1995 through 2002, the pre–Reading First period. Because third grade was the first tested grade in Texas during this period, we do not observe a lagged own-subject reading score for third graders. We address this limitation by constructing a value-added measure that consists of regressing standardized third-grade reading scores on lagged attendance (e.g., second-grade attendance), a student's third-grade math score, and student demographics (indicators for race, gender, economic disadvantage, ESL status, gifted status, special education, and immigrant status).[19] We take the residuals from the first-stage regression and collapse them to campus–year means. These means serve as our raw campus value-added estimates. We apply a standard empirical Bayes shrinkage procedure to the value added estimates, but the results are not meaningfully different if the raw value added estimates are used instead (Kane and Staiger, 2008).

To examine whether Reading First shifted the composition of the teaching workforce, we assign each teacher a value-added score on the basis of the campus where they were employed in 2002, the year used to construct the Reading First eligibility criteria. Thus, every teacher at a given campus receives the same score. For each campus and year, we then average the 2002 scores of the teachers working there. If there were no teacher turnover or movement, these averages would not change. We construct these means for all classroom employees at a campus and separately for teachers who entered, exited, or remained at the campus (incumbents). Figure A13 reports the results. Across all models, we find no evidence that teachers with higher value added moved into Reading First districts after the program began. If anything, the estimates point to slightly lower average pre-2003 value added in Reading First districts, although value added is a noisy measure. These findings suggest that the reading achievement effects we document were not driven by a reshuffling of higher-quality teachers across Reading First and non–Reading First 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 A14. The instrumental variables estimates imply that Reading First funding lowers the

[19] Figure A13 also considers two alternatives. One where we do not include the third-grade math score and another measure of teacher value added that regresses standardized fourth-grade reading scores on fourth-grade student demographics, third-grade attendance, and the student's third-grade reading score. In all cases, our results are similar.

no-degree share and raises the bachelor's share by roughly the same amount among employees entering and leaving the classroom. The difference-in-differences estimates are similar in sign but imprecise.

We next examine whether Reading First affected teacher staffing levels. Figure A15 reports estimates where the outcomes are various measures of full-time-equivalent (FTE) teachers per student. We consider the overall teacher-to-student ratio, the rates of new and departing teachers, K–3 FTE teachers per student, and K–3 reading/English Language Arts (ELA) FTE teachers per student. Overall, we find no consistent effect of Reading First on the number of teachers per student. The difference-in-differences estimates are small and not statistically significant across all these staffing measures. However, the instrumental variables estimates suggest a decline in K–3 reading/ELA FTE teachers per student. Most of the decline comes from incumbent employees, who stayed on the same campus and whose total K–3 FTE does not change. This pattern suggests a change in how these employees' positions were assigned or classified rather than a loss of staff.

Figure A16 examines teacher compensation, estimating the effect on salary per FTE teacher (in thousands of dollars). We cannot observe which schools or employees were allocated Reading First funds since Reading First funding is observed only in the district-level finance data and not in this employee-level dataset. Because Reading First funds were distributed across a subset of campuses within a district and likely reached only a subset of staff within those campuses, this analysis yields estimates that are attenuated. Even so, the instrumental variables estimates imply that Reading First funding raised salary per FTE for all classroom employees. The difference-in-differences estimates are mostly positive but imprecise. This increase is consistent with our earlier finding that districts spent Reading First dollars partly on salaries, wages, and benefits.

Taken together, we find no evidence that Reading First raised reading scores by adding teachers, attracting teachers with higher value added, or changing the degree mix of classroom employees. The instrumental variables estimates do suggest that districts receiving more Reading First money paid their staff more and had a larger share of employees with bachelor's degrees. Because the difference-in-differences estimates are similar in sign but imprecise, we view this pattern as suggestive. The instrumental variables estimates place the most weight on the districts that received the most funding, which are also the districts that we find had the smallest reading gains per dollar (Appendix Figure A4). One interpretation is that where funding was largest, districts invested a larger share of it in staff pay and a smaller share in reading materials. Higher pay may still benefit students in other ways, for example, by helping schools retain teachers.

5.4 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. The way awarded

districts spent Reading First funds was not unrestricted. Instead, funds were tied to curricular mandates, coaching requirements, and professional development, all of which could have contributed to improvements in educational gains alongside the money itself. Given that a large literature finds that additional school resources raise achievement and attainment (Jackson et al., 2016; Lafortune et al., 2018; Abott et al., 2020; Baron, 2022; Biasi, 2023; Jackson and Mackevicius, 2024), we take the value of money as given in Texas and ask whether a dollar that arrived with instructions on its use did more than a dollar that did not. To do so, we compare the effects of Reading First with those of a more general revenue shock that has been widely used in the school spending literature, the passage of a school district tax election (Cellini et al., 2010; Martorell et al., 2016; Abott et al., 2020; Baron, 2022; Biasi et al., 2025). We follow past work and leverage Texas school districts' ability to petition voters to raise the operational tax rate. In contrast to Reading First grants, these funds carry no programmatic requirements and can be spent at the district's discretion. If the Reading First results reflected additional resources alone, independent of curricular changes, we would expect a tax election windfall of equivalent magnitude to produce similar reading gains as those of Reading First. Given that we also observe what districts spent election revenue on, we can compare the two sources of funds on how they were spent as well as on what they produced.

We estimate a difference-in-differences model using data on Texas school district tax elections between 2000 and 2019 merged with the Texas PEIMS financial data discussed earlier.[20] Treated districts are those where a tax increase passed and control districts those that had not yet passed one or never did. We use the year of the first successful tax increase as the treatment year. Figure C1 presents the event study for local per student revenue following a successful tax election. Local revenue rises after passage by roughly $321 per student on average over the post-election years. This figure also shows a rise of similar size in federal revenue per student in eligible districts as a result of Reading First. Table C1 examines how these discretionary dollars were spent. As expected, school revenue increased for the average district after a successful tax election, with most of that increase coming from local revenue sources. The remaining revenue reflects state aid since many of the later elections involved rate swaps. Districts directed the new revenue predominantly to salaries and wages, with no significant change in supplies and materials. When we decompose the expenditure effect further into the supplies subcategories, we find no increase in spending on reading materials or general supplies, the categories that account for a substantial portion of Reading First spending. Spending on salaries is a productive use of money in its own right. Although the two sources of funds resulted in different inputs, we cannot attribute the difference in their effects on reading scores to any particular one. However, this comparison suggests that, in the same districts and years, a dollar that came with Reading First's curricular requirements raised reading scores by more than a dollar that districts allocated on their own. This evidence is consistent with the literacy components

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

attached to Reading First having contributed to the reading gains we find.

Panel C of Figure C1 shows the effects of tax elections and of Reading First on third-grade reading scores. Both point estimates capture the average treatment effect across the 2000–2019 period and come from the same campus–year panel with campus and year fixed effects. The passage of a tax increase raised reading scores by 0.02 standard deviations. The estimate is imprecise, and we do not interpret this finding as evidence that unrestricted spending has no effect on test scores. Jackson and Mackevicius (2024) find that a $1,282 (in 2025 $) increase in per-pupil spending sustained over four years raises test scores by about 0.03 standard deviations on average, which implies a gain of roughly 0.01 standard deviations for the $428 increase in total revenue per student that followed a successful election (Table C1). Our estimates are in line with this benchmark, and the literature finds that such effects build gradually over several years (Lafortune et al., 2018; Abott et al., 2020). Conversely, Figure C1 shows that Reading First, which brought a comparable amount of revenue per student to eligible districts, raised scores by roughly 0.12 standard deviations.

6 Long-Run Effects

Thus far, we have found that Reading First improved student test scores, in a way that durably changed classroom instruction. This section asks whether the resulting achievement gains persisted as exposed students aged, first through grades four to eight and then into high school, college, and the labor market. As described in Section 2, the ERC data link students to high school completion records, postsecondary enrollment and degree completion from the Texas Higher Education Coordinating Board, the National Student Clearinghouse, and quarterly earnings from the Texas Workforce Commission. We estimate fixed effect ordinary least squares models for test scores, high school and college outcomes and fixed effect Poisson models for employment and earnings. Figure 6 and Table B12 report the results from our difference-in-differences specification. The first column of Figure 6 (Panels A–C) examines third through twelfth grade outcomes, and the second column (Panels D–F) examines college and labor market outcomes. Each plotted coefficient comes from a separate specification, and for every outcome, we report estimates for the full sample of districts and for districts within five percentage points of the eligibility threshold. Analogous instrumental variables estimates are shown in Appendix Figure A17.

6.1 Persistence Through High School

We first examine whether the Reading First effect on reading test scores persisted beyond third grade for an individual student. Panel A of Figure 6 reports our main difference-in-differences estimate for third-grade reading at the student rather than the campus–year level, still including campus and year fixed effects. The estimates are similar to the more aggregated results presented earlier, with eligibility increasing third-grade reading scores by roughly 0.13

standard deviations. Panel A also plots effects on reading scores through grade eight. The gains persisted across grades, waning over time but remaining statistically and economically meaningful for all grades. Panel B repeats this exercise for math. The individual-level estimates are again similar to our earlier campus-level results. Panel C of Figure 6 reports the effects of Reading First on whether a student ever dropped out, received a standard diploma on time (within four years), and graduated from high school. Reading First students were 1.4 percentage points less likely to drop out, 2.7 percentage points more likely to receive a standard diploma on time, and 2.3 percentage points more likely to ever graduate from high school.

These high school outcomes use third-grade cohorts through 2014, the last cohort to have reached twelfth grade by 2023, the final year of the enrollment records, so that no student is counted as a nongraduate before having had the chance to graduate. The administrative data contain detailed information about students who dropped out of school beginning in seventh grade. Similarly, there is detailed information about who graduated from high school and when and the manner in which they graduated (i.e., with an on-time diploma or a later GED certificate). But to measure dropout and graduation rates, we also need accurate information on the denominator, which is not as simple to determine. We cannot simply divide high school graduates in a given year by the number of third graders nine years before. We also cannot just look at the likelihood that a given third grader shows up as a high school graduate later on. This is because there are many legitimate reasons why a student might leave the dataset. They could have moved out of state, gone to a private school, become incarcerated, or even died. An additional limitation is that we cannot observe dropouts before seventh grade.

Therefore, in the high school dropout, high school completion, and on-time completion analyses, we limit the sample to students whose final school status we can observe. For graduation and on-time completion, these are either students with a graduation record, students enrolled in twelfth grade who never graduated, or students with an official dropout record prior to twelfth grade. A student who left Texas public schools before twelfth grade without a dropout record is excluded rather than counted as a nongraduate since this could be a legitimate matriculation. For the sample of students considered in the dropout analysis, we include every student observed in seventh grade or later, the grade from which Texas begins reporting dropout status.

These types of denominator changes raise attenuation concerns. Since this setting essentially corresponds to a type of intent-to-treat difference-in-differences design, the fact that some students considered treated here may have exited the sample, might attenuate our treatment effect estimates. Thus, in the appendix, we report a robustness check where we restrict to the set of students whom we observe continuously from third grade until their "final school" status can be credibly estimated. Since this begins in the seventh grade, if Reading First affected a student's likelihood of remaining in the sample, this criterion might introduce endogeneity, but it comes at the benefit of keeping the observed sample of students consistent across time. Given the endogeneity concern, this is not our preferred specification, only a robustness check. We report results for this set of students in Figure A18, where we find smaller standard errors

and higher magnitude of the effects.

6.2 College and the Labor Market

We next examine the effects of Reading First on college and labor market outcomes. These outcomes require that we observe students into adulthood, and a student who is too young to have achieved an outcome should not be counted as not having attained it. We therefore restrict college outcomes to the cohorts old enough to be observed to a given age. The enrollment outcomes use third-grade cohorts through 2014, the last cohort observed to age 19 in the enrollment records, and the degree outcomes use cohorts through 2011, the last cohort observed to age 22. Within these cohorts, we report two definitions: whether the outcome is ever observed in our data and whether it occurred by the stated age; the latter holds the observation window fixed across cohorts. The employment and wage estimates use cohorts through 2014.

Panel D of Figure 6 shows that Reading First increased the probability of a student ever enrolling in college by about 1.3 percentage points, although this estimate is imprecise, and the probability of enrolling by age 19 by 1.9 percentage points. Reading First also increased the probability of a student ever completing a college degree by 0.8 percentage points. This gain is concentrated among four-year institutions. Four-year degree completion rises by 1.0 percentage point, but two-year degree completion falls by 0.5 percentage points. One interpretation is that improved literacy matters most at the four-year degree margin. Two-year colleges typically have less restrictive admissions than four-year institutions, so an intervention that improves academic skills may matter more for admission to the latter. The negative estimate for two-year degrees may also indicate substitution, since students who would otherwise have attended a two-year college might instead have enrolled in a four-year institution, a shift that would also lead to an increase in earnings (Mountjoy, 2022). These college estimates are less precise than the high school results, and the degree estimates defined by age 22 are smaller, reflecting that many degrees are completed after that age.

Panels E and F of Figure 6 examine students' likelihood of being in the labor force and associated annual wages at each age from 18 to 25.[21] For these outcomes, we estimate fixed effect Poisson models. One complication in the wage data is the interpretation of zeros, which we address by imputing zero earnings for any individual who appears in the wage data at least once. The results in Figure 6 suggest that Reading First exposure raised earnings in adulthood. The likelihood of being employed is concentrated at age 18, fades by age 19, and is absent at later ages. Wages, by contrast, are higher at every age from 18 to 25, with the most precise and statistically significant estimates for ages 24 and 25—an effect timing that coincides with the college graduation effects. The difference-in-differences estimate in Appendix Table B12 implies that Reading First eligibility increased annual wages at age 25 by $509, or approximately 1.5% of the $33,713 mean among untreated students. The Poisson estimate plotted in the figure,

[21] The earnings sample includes students who turned 25 between 2010 and 2025, the years covered by the available wage records.

0.019 log points, is of a similar size.[22] These magnitudes are consistent with other estimates of how early educational inputs determine adult earnings. The effect of eligibility is comparable to that of assignment of every student to a one-standard-deviation-better teacher for a single year (Chetty et al., 2014b).

In the instrumental variables specification (Appendix Figure A17), the high school and college effects are close to the difference-in-differences estimates, and both sets of estimates are robust to restricting the sample to continuously enrolled students (Appendix Figures A18 and A19).[23] However, the employment estimates are slightly more negative and more precise, and the age-25 wage estimate is positive but smaller and imprecise. Examining the effects across the amount of funding districts received, helps shed light on this discrepancy. Panels C and D of Appendix Figure A4 report the wage effects at age 25 separately for eligible districts in three groups of cumulative predicted funding (under $5 million, $5 million to $15 million, and $15 million or more). Panel C evaluates each group's effect at its own mean funding, and Panel D rescales every estimate to the same $20 million. In levels, the wage effect rises from the low-spending group to the medium-spending group and then falls. Students in the highest-funded districts gained less than students in districts that received a fraction of the money. Per dollar, the wage effect declines from the low-spending group to the high-spending group, with the low-spending group's return roughly eight times that of the high-spending group. As with reading, most of the variation in predicted funding comes from a handful of very large districts, so the pooled per dollar estimate is effectively the high-spending group's estimate. Since that group has the smallest per dollar wage effect, the pooled estimate understates what the program did for wages in the low- and medium-spending districts. This is also where wages and reading differ. In Appendix Figure A4, the high-spending group's reading gains (Panel A) match the medium-spending group's, but its wage gains (Panel C) fall short of them. This pattern is consistent with the wage effects being driven more by participation in the program than by the amount of money districts received.

Given those wage gains, it is natural to ask why employment measured at later ages does not respond and why the instrumental variables estimate is, if anything, slightly negative. One possibility follows from the college results. Because Reading First increased four-year college graduation, some treated students may have enrolled in colleges out of state and remained outside Texas while employed or have graduated from a Texas university and taken a job in another state. In either case, coding these students as having zero earnings would attenuate our estimates.

[22] We note two limitations of the age-25 earnings measure: 1) it captures only a fraction of the lifetime returns since it falls in the early stage of someone's professional career, and 2) students who pursue graduate education earn little while still enrolled. Given our finding that Reading First increased college enrollment, treated students may still have been in school at age 25. Changes on this margin would push the estimate toward zero, and the $509 figure likely understates the long-run impact of Reading First.

[23] A student is continuously enrolled if we observe a high school outcome for them, meaning they graduated, reached twelfth grade, or appear in the dropout files. Students who left the public school system before twelfth grade with no dropout record are excluded. This removes students who moved out of state or to private schools, whose later schooling and earnings our records would understate.

To understand the channels behind the earnings gain attributed to Reading First, Figure 7 presents a Gelbach decomposition (Gelbach, 2016), which simultaneously decomposes the total wage effect into the portions statistically accounted for by intermediate outcomes observed in the data. This type of accounting does not depend on the order in which the channels are entered and should be interpreted as a descriptive approach of exploring the incidence of the wage effect, rather than as indicative of a set of independent causal mechanisms. An on-time high school diploma accounts for roughly $140 of the total effect, enrollment in and completion of college (institution fixed effects) account for $303, field of study (CIP fixed effects) for $114, and industry of employment (NAICS fixed effects) for $163. The component unexplained by these channels is negative (−$211). This pattern suggests that a majority of the long-run earnings effect is attributable to improved educational attainment and the resulting occupational upgrading, rather than to unobserved channels. Appendix Table B13 reaches the same conclusion using expected-wage measures that assign each student the control group mean wage for their college, field of study, or industry. This finding aligns with the literature on returns to higher education that identifies a wage premium from attending a public four-year university (Black et al., 2023; Denning et al., 2019; Kozakowski, 2023; Mountjoy, 2026; Smith et al., 2025; Zimmerman, 2014). Both Kozakowski (2023) and Mountjoy (2026) show that the wage premium appears eight years after a student applies to college, approximately at ages 24–25, and grows over time. Our estimates of a wage premium are small in magnitude relative to others in the literature since we observe Reading First treated students only in the very early stages of their professional career (through age 25).

Interestingly, the decomposition reveals that the smallest contributor is field of study. One hypothesis is that reading improvements that persist through grade school may lead some to choose more reading-dependent fields. In Appendix Table C2, we examine this possibility directly. The Occupational Information Network (O*NET) is a Department of Labor database that scores occupations on the importance of specific skills, including reading comprehension. Using these reading importance scores, we examine the extent to which Reading First exposure altered the reading intensity of chosen majors, colleges, and broad industry of occupation (at age 25). We find no evidence that Reading First exposure shifted people into majors, colleges, or industries where reading is more important.

7 Value of Reading First per Dollar Spent

Reading First was inexpensive relative to its effects. In fiscal terms, eligible districts received roughly $75 per K–3 student per year over the full post-period, with an associated treatment effect of 0.11 standard deviations in reading score improvement. During the program (2004–2011), the $132 per student–year produced an effect of 0.07 standard deviations. Note that the full post-period estimate benefits from persistence, which required no additional spending. Under either estimate, the implied cost per 0.1 standard deviations is between $70 and $184 per student–year, compared to the roughly $3,846 (in 2025 $) per pupil per year that the

school spending literature implies is typically required for a comparable gain (Jackson and Mackevicius, 2024).[24]

To formalize the comparison, we compute the Marginal Value of Public Funds (MVPF) of Reading First, following Hendren and Sprung-Keyser (2020) and using the difference-in-differences estimates rather than the pooled instrumental variables slope. The MVPF is the ratio of beneficiaries' willingness to pay to the net cost borne by the government, where net cost is the program spending less the additional tax revenue generated by the resulting increase in earnings. We adopt the analysis of Jackson et al. (2016) for K–12 school spending effects and use an annual earnings gain of \$509 at age 25 against cumulative Reading First during-program spending (assuming three years of exposure, discussed in more detail in Appendix C.3) of roughly \$397 per exposed student, in 2025 dollars. When assessed relative to baseline earnings of \$33,713, the \$509 Reading First–induced wage premium represents a 1.5% increase, or 0.0038% per program dollar spent. We hold this earnings gain constant throughout the entire age-earnings profile, allow for 0.5% annual real wage growth, and project the gain forward through age 65. We discount both benefits and costs back to the average age at which the money was spent and Reading First students were exposed to the treatment (age 7).

Under these assumptions, the additional tax revenue exceeded the initial outlay—meaning the program paid for itself, so the MVPF is unbounded—once the treated students had reached age 31. Appendix C.3 sets out the construction in full. Table C3 stops the earnings projection at successively later ages and reports the net cost to the government at each. Crediting Reading First with only the single year of earnings we observe, at age 25, and assuming the effect vanished thereafter, the government recovered fourteen cents of each dollar spent, and the MVPF is 0.71. We note that this projection is conservative given that our method holds the proportional earnings effect constant, whereas Mountjoy (2026) finds that the earnings returns to attending a Texas public university more than double between ages 25 and 30.

8 Conclusion

Reading proficiency is an important predictor of later-life achievement. Although substantial resources have been allocated toward improving reading curricula in elementary schools, evidence of the long-term benefits is limited. We study a phonics-based policy intervention that improved reading scores among third graders in Texas, sustained those improvements through later grades, and had performance spillovers in math. Using administrative data from the TEA 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. In contrast to influxes of unconditional funding, Reading First included specific guidelines on how districts should

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

allocate the money they received. We track those expenditures and show that they went to a large extent to reading materials and teacher salaries. At a direct fiscal cost of roughly $70–$184 per student–year, Reading First achieved an increase of 0.1 standard deviations in reading test scores, making it more cost-effective than other interventions studied in the literature (Jackson and Mackevicius, 2024; Guryan et al., 2023).

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 on longer, more complex questions than their ineligible peers. Following treated students through college and the early stages of their labor market participation, we find an increased probability that students enrolled in college and graduated from four-year universities and earned a wage premium of $509 at age 25. Overall, our work informs the debate over the effectiveness of early childhood interventions by showing that their returns depend on how government funding is spent, not only on how much is allocated. At relatively low cost, Reading First generated lasting effects that substantially improved both short- and long-term outcomes for exposed students.

References

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

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

Ainsworth, Robert, Rajeev H. Dehejia, Andrei Munteanu, Cristian Pop-Eleches, and Miguel Urquiola (2025) "Malleable Minds: The Effects of STEM v. Humanities-Focused Curricula," NBER Working Paper34502, National Bureau of Economic Research, https://www.nber.org/papers/w34502.

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

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

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

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

Andersen, Christina M., Jørgen Brandt, Jesper H. Christensen, Lise M. Frohn, Camilla Geels, Timo Hener, Marianne Simonsen, and Lars Skipper (2025) "Air Pollution and Cognition in Children," *Journal of Human Resources*, 10.3368/jhr.1024-13871R1.

Andersen, Erik T. J., Simon Graffy, Jason T. Kerwin, and Monica P. Lambon-Quayefio (2026) "How to Build a Reader: Evidence from a Scalable Literacy Intervention in Ghana," Working paper.

Anderson, Michael L., Justin Gallagher, and Elizabeth Ramirez Ritchie (2018) "School meal quality and academic performance," *Journal of Public Economics*, 168, 81–93, 10.1016/j.jpubeco.2018.09.013.

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

Arold, Benjamin W (2024) "Evolution vs. Creationism in the Classroom: The Lasting Effects of Science Education," *Quarterly Journal of Economics*, 139 (4), 2331–2375, 10.1093/qje/qjae019.

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

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

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

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

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden (2007) "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics*, 122 (3), 1235–1264, 10.1162/qjec.122.3.1235.

Baron, E Jason (2022) "School spending and student outcomes: Evidence from revenue limit

elections in Wisconsin," *American Economic Journal: Economic Policy*, 14 (1), 1–39.

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

——— (2023) "School Finance Equalization Increases Intergenerational Mobility," *Journal of Labor Economics*, 41 (1), 1–38.

Biasi, Barbara, Julien Lafortune, and David Schönholzer (2025) "What Works and for Whom? Effectiveness and Efficiency of School Capital Investments across the U.S.," *Quarterly Journal of Economics*, 140 (3), 2329–2379.

Black, Sandra E., Jeffrey T. Denning, and Jesse Rothstein (2023) "Winners and Losers? The Effect of Gaining and Losing Access to Selective Colleges on Education and Labor Market Outcomes," *American Economic Journal: Applied Economics*, 15 (1), 26–67.

Blazar, David, Blake Heller, Thomas J. Kane et al. (2020) "Curriculum Reform in the Common Core Era: Evaluating Elementary Math Textbooks across Six U.S. States," *Journal of Policy Analysis and Management*, 39 (4), 966–1019, 10.1002/pam.22257.

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

Brunner, Eric, Joshua Hyman, and Andrew Ju (2020) "School Finance Reforms, Teachers' Unions, and the Allocation of School Resources," *Review of Economics and Statistics*, 102 (3), 473–489, 10.1162/rest_a_00828.

Buhl-Wiggers, Julie, Jason T. Kerwin, Ricardo Montero de la Piedra, Jeffrey A. Smith, and Rebecca Thornton (2024) "Reading for Life: Lasting Impacts of a Literacy Intervention in Uganda," Working paper.

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

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

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

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

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein (2010) "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design," *Quarterly Journal of Economics*, 125 (1), 215–261.

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

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

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

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

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

Clemens, Jeffrey, Philip G. Hoxie, and Stan Veuger (2026) "Intergovernmental Grants to School Districts and Educational Outcomes during the COVID-19 Pandemic," NBER Working Paper35447, National Bureau of Economic Research, 10.3386/w35447.

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

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

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

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

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

Curry, Janice (2005) "AISD K–5 Accelerated Reading and Mathematics Instruction Evaluation, 2004–2005," Technical ReportPublication 04.07, Austin Independent School District, Department of Program Evaluation, Austin, TX.

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

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

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

Denning, Jeffrey T., Benjamin M. Marx, and Lesley J. Turner (2019) "ProPelled: The Effects of Grants on Graduation, Earnings, and Welfare," *American Economic Journal: Applied Economics*, 11 (3), 193–224.

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

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

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

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

Ellison, Glenn and Parag A. Pathak (2025) "Optimal School System and Curriculum Design: Theory and Evidence," NBER Working Paper34091, National Bureau of Economic Research, https://www.nber.org/papers/w34091.

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

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

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

García, Jorge Luis, James J. Heckman, and Victor Ronda (2023) "The Lasting Effects of Early-Childhood Education on Promoting the Skills and Social Mobility of Disadvantaged African Americans and Their Children," *Journal of Political Economy*, 131 (6), 1477–1506,

10.1086/722936.

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

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

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

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

Gray-Lobe, Guthrie, Parag A Pathak, and Christopher R Walters (2023) "The Long-Term Effects of Universal Preschool in Boston," *Quarterly Journal of Economics*, 138 (1), 363–411, 10.1093/qje/qjac036.

Gruber, Jon and Emmanuel Saez (2002) "The Elasticity of Taxable Income: Evidence and Implications," *Journal of Public Economics*, 84 (1), 1–32.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt et al. (2023) "Not Too Late: Improving Academic Outcomes among Adolescents," *American Economic Review*, 113 (3), 738–765, 10.1257/aer.20210434.

Handel, Danielle V. and Eric A. Hanushek (2023) "U.S. School Finance: Resources and Outcomes," in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann eds. *Handbook of the Economics of Education*, Vol. 7, 143–226, Amsterdam: North-Holland.

Hanford, Emily (2020) "What the Words Say," APM Reports, August, https://www.apmreports.org/episode/2020/08/06/what-the-words-say.

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

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

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

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

Hollingsworth, Alex, Jiafang Mike Huang, Ivan Rudik, and Nicholas J. Sanders (2022) "A Thousand Cuts: Cumulative Lead Exposure Reduces Academic Achievement," *Journal of Human Resources*, 10.3368/jhr.0222-12169R2.

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

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

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

Jackson, C. Kirabo (2018) "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes," *Journal of Political Economy*, 126 (5), 2072–2107, 10.1086/699018.

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

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

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

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

John, Anett, Malak Koura, and Todd Pugatch (2026) "Improving Literacy Through Cognitive Science: Evidence from the United Arab Emirates," J-PAL evaluation summary, https://www.povertyactionlab.org/evaluation/improving-literacy-through-cognitive-science-evidence-united-arab-emirates, Working paper in progress.

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

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

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

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

Kozakowski, Whitney (2023) "Are Four-Year Public Colleges Engines for Economic Mobility? Evidence from Statewide Admissions Thresholds. EdWorkingPaper No. 23-727.," *Annenberg Institute for School Reform at Brown University*.

Kreisman, Daniel and Matthew P. Steinberg (2019) "The Effect of Increased Funding on Student Achievement: Evidence from Texas's Small District Adjustment," *Journal of Public Economics*, 176, 118–141, 10.1016/j.jpubeco.2019.04.003.

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

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach (2018) "School Finance Reform and the Distribution of Student Achievement," *American Economic Journal: Applied Economics*, 10 (2), 1–26.

Lee, Steven and Matthew Schaelling (2026) "Content Relatability and Standardized Testing: Evidence from Texas," Working paper.

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

Loveless, Tom (2021) *Between the State and the Schoolhouse: Understanding the Failure of Common Core*, Cambridge, MA: Harvard Education Press.

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

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

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

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

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

Menon, Shailaja and Elfrieda H. Hiebert (1999) "Literature Anthologies: The Task for First-Grade Readers," CIERA Report1-009, Center for the Improvement of Early Reading Achievement, University of Michigan, Ann Arbor, MI, https://files.eric.ed.gov/fulltext/ED436754.pdf.

Moreno-Medina, Jonathan, Eric Nielsen, and Viviana Rodriguez (2026) "The Anatomy of a High-Price Question: Text, Skills, and the Economics of Achievement Measurement," EdWorkingPaper26-1467, Annenberg Institute at Brown University, 10.26300/jfph-7511.

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

Moss, Marc, Alyssa Rulf Fountain, Beth Boulay, Megan Horst, Chris Rodger, and Melanie Brown-Lyons (2008) "Reading First Implementation Evaluation: Final Report," technical report, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, U.S. Department of Education, Washington, DC.

Mountjoy, Jack (2022) "Community Colleges and Upward Mobility," *American Economic Review*, 112 (8), 2580–2630, 10.1257/aer.20181756.

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

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

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

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

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

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

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

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

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

Polikoff, Morgan S. (2021) *Beyond Standards: The Fragmentation of Education Governance and the Promise of Curriculum Reform*, Cambridge, MA: Harvard Education Press.

Rayner, Keith (1998) "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, 124 (3), 372–422.

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

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

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

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

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

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

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

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

Texas Education Agency (2003a) "Request for Application #701-03-018: Texas Reading First Initiative (TRFI)," technical report, Texas Education Agency, Austin, TX, Cycle 1 subgrant competition. Published in the Texas Register May 16, 2003; applications due July 10, 2003.

——— (2003b) "State of Texas: Reading First Application."

——— (2004) "Request for Application #701-04-016: Texas Reading First Initiative for Grades K–3, Cycle 2," technical report, Texas Education Agency, Austin, TX, Cycle 2 subgrant competition. Published in the Texas Register March 26, 2004; applications due June 3, 2004.

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

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

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

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

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

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

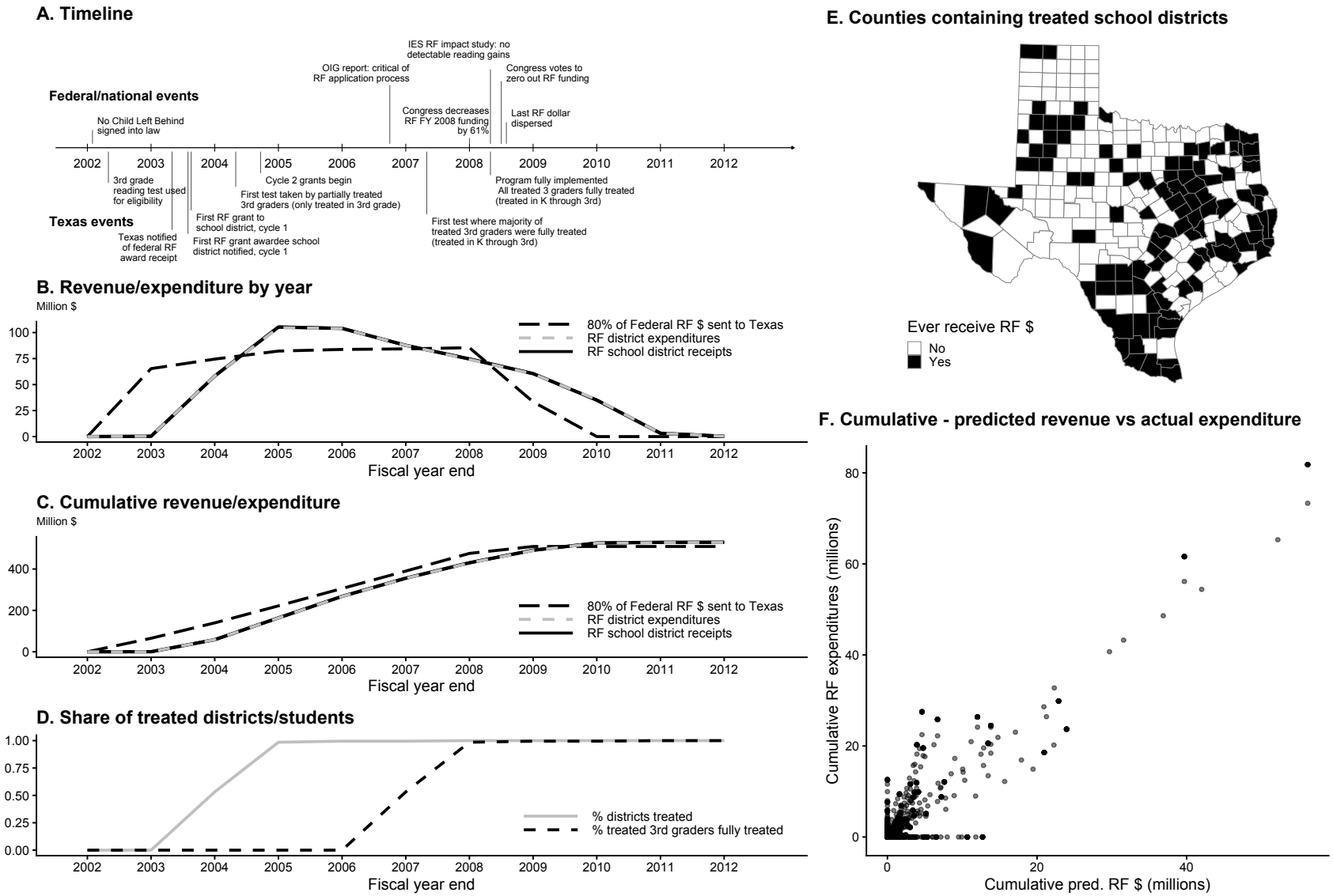
Xu, Zeyu, Umut Özek, Jesse Levin, and Dong Hoon Lee (2023) "Effects of Large-Scale Early Math Interventions on Student Outcomes: Evidence From Kentucky's Math Achievement Fund," *Educational Evaluation and Policy Analysis*, 01623737241303142.

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

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

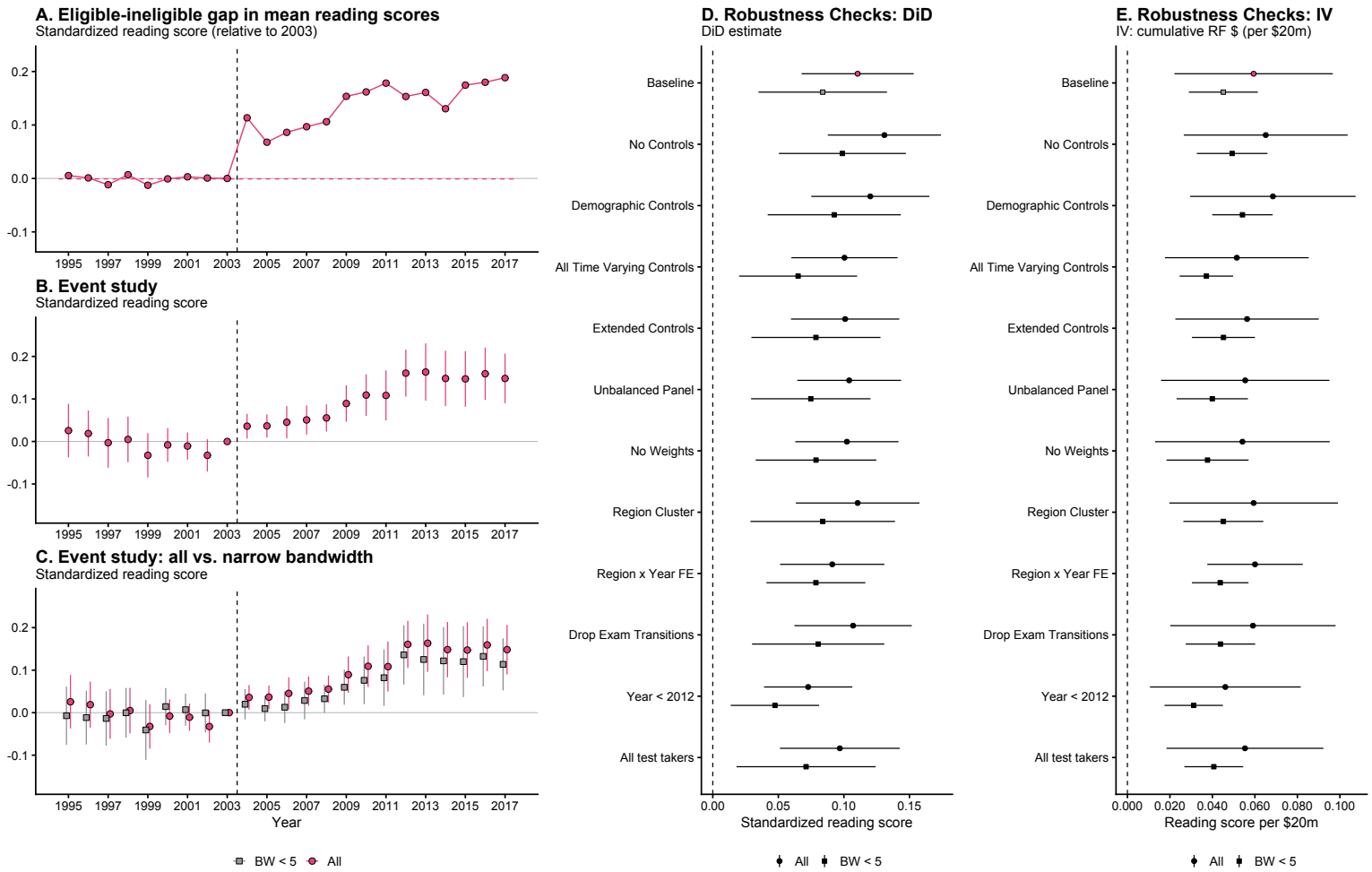
Zimmerman, Seth D. (2014) "The Returns to College Admission for Academically Marginal Students," *Journal of Labor Economics*, 32 (4), 711–754, 10.1086/676661.

Figure 1: Overview of spatial and temporal sources of variation



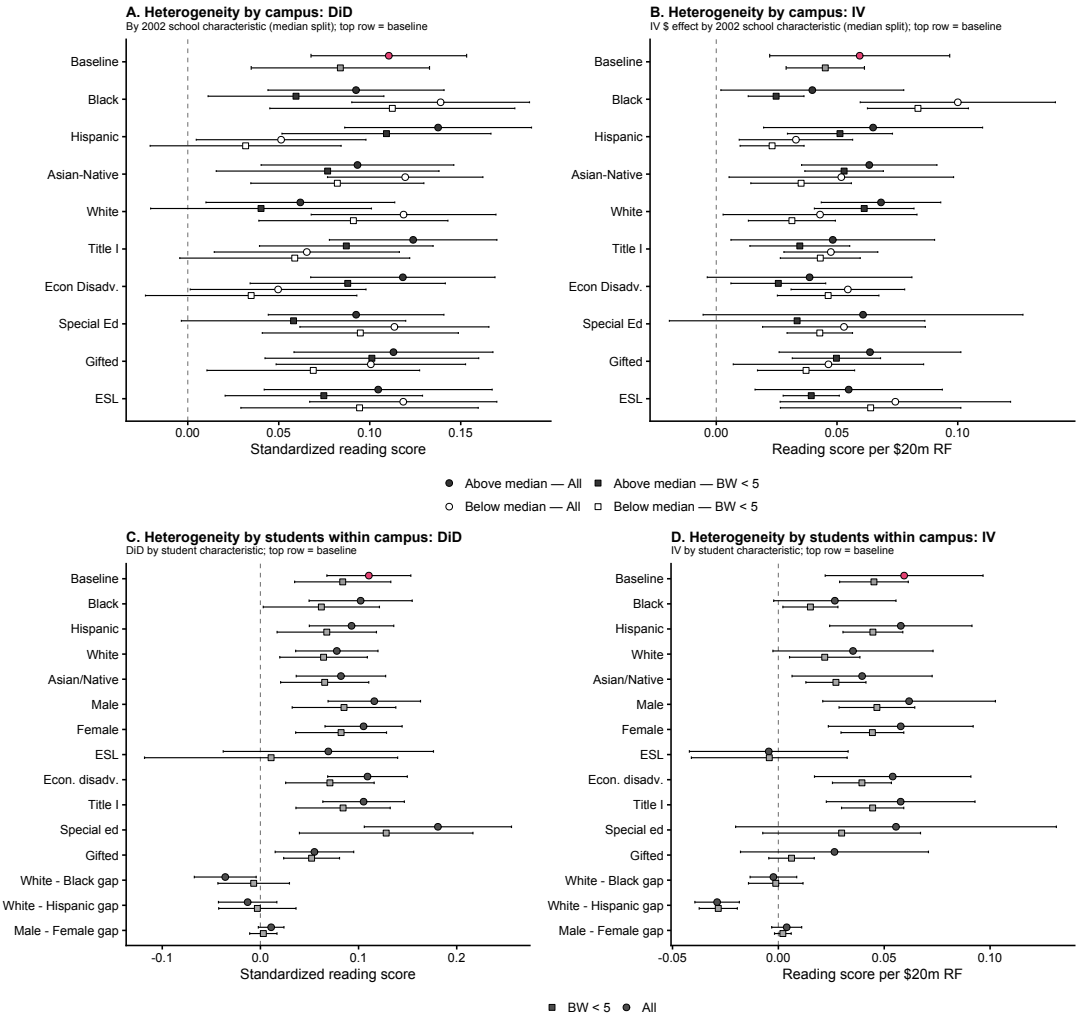
Notes: Panel A presents a timeline of Reading First (RF) in Texas alongside key federal events. Panel B (C) plots annual (cumulative) district RF receipts and expenditures by year in millions of dollars. For every federal dollar received, Texas had to distribute at least 80% to school districts. Panel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

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



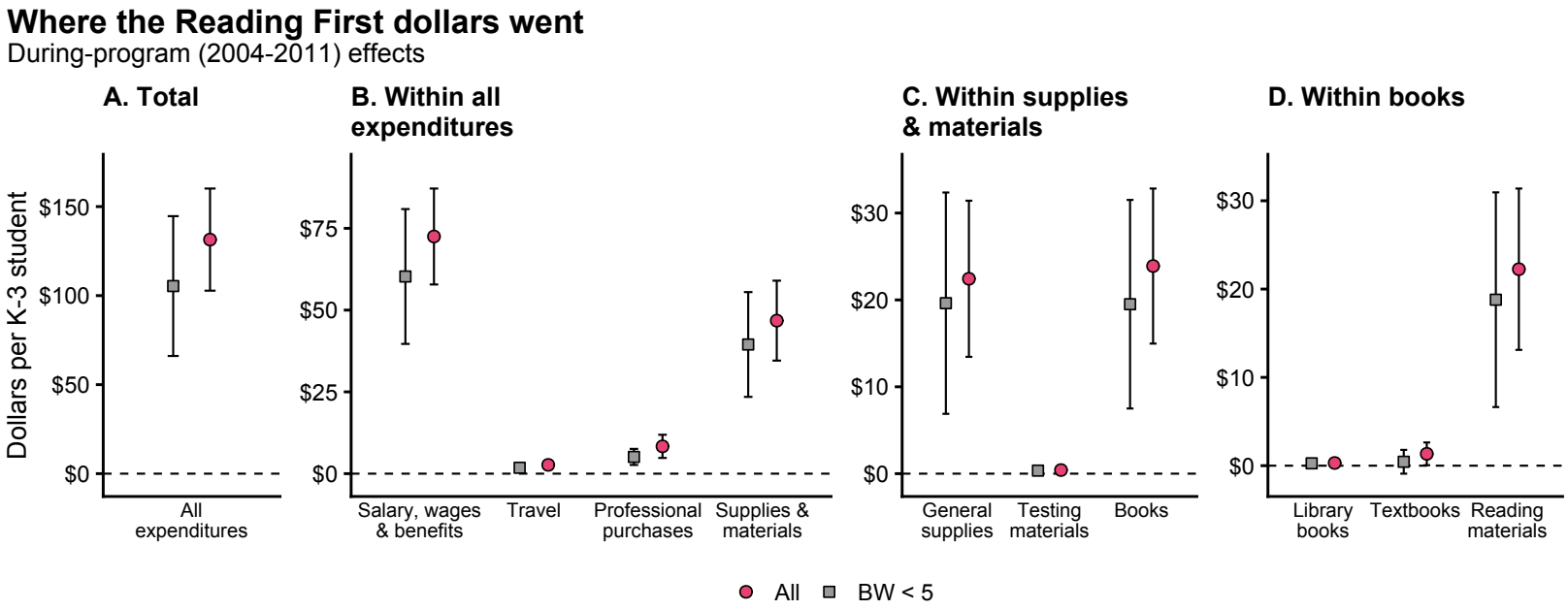
Notes: Panel A plots the raw difference in standardized third-grade reading scores between districts eligible and ineligible for Reading First, relative to 2003. Panel B presents event study estimates from equation (3) with 95% confidence intervals. Panel C adds a restricted sample of districts within five percentage points of the 2002 statewide average of 13.6% (square markers). Panel D reports specification checks for the difference-in-differences model, and Panel E reports the corresponding checks for the instrumental variables model, each for the full and restricted samples (see Section 4.2 for details). All models include campus-by-grade and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects. Standard errors clustered at the district level. All samples are restricted to first-time test takers, except for the bottom row of Panels D and E, which uses all test takers. Panel E estimates are per $20 million of cumulative Reading First funding.

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



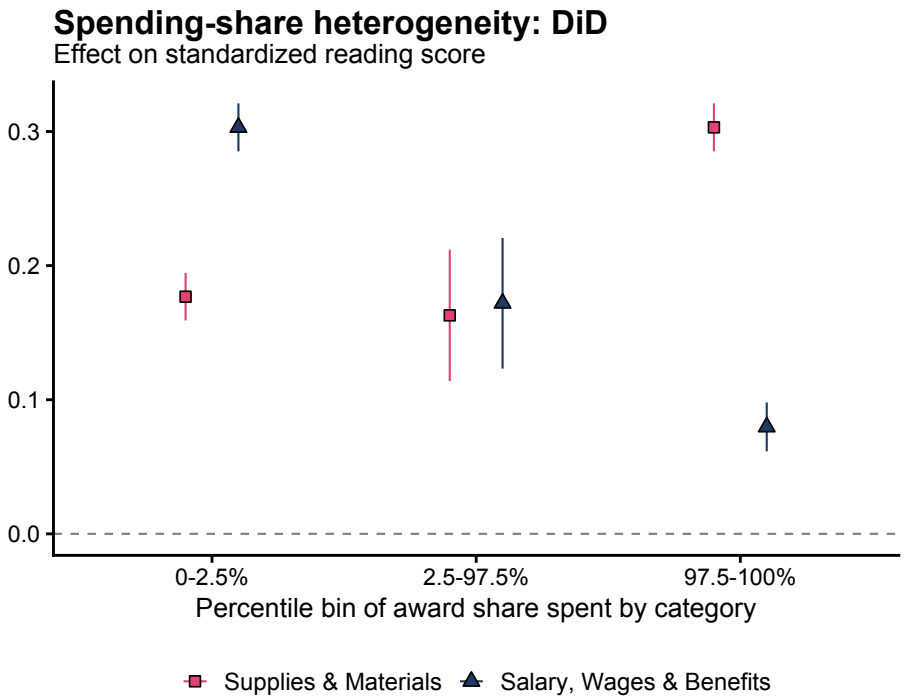
Notes: Panels A and B report split-sample estimates by 2002 campus characteristics (above/below median). Panel A reports difference-in-differences estimates, and Panel B reports instrumental variables estimates per $20 million of cumulative Reading First funding. Panels C and D report estimates by student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

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



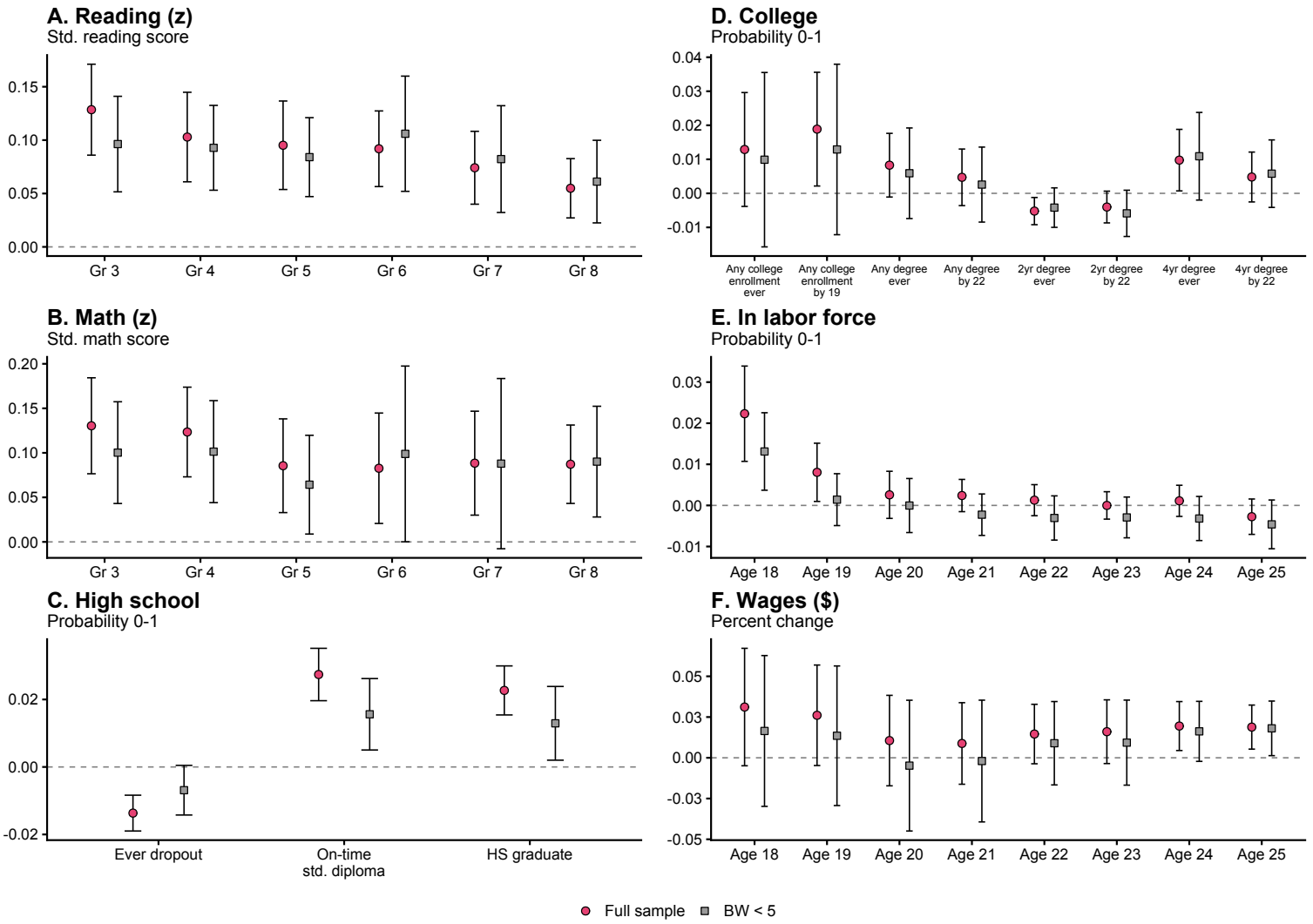
Notes: Each point reports the estimated effect of Reading First eligibility on the indicated Reading First expenditure category per K–3 student during the program period (2004–2011). Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. The four panels read left to right as a hierarchy, and the last category of each panel is decomposed in the next. Panel A reports total expenditures. Panel B splits total expenditures into salary, wages & benefits; travel; professional purchases; and supplies & materials. Panel C splits supplies & materials into general supplies, testing materials, and books. Panel D splits books into library books, textbooks, and reading materials. Each panel has its own dollar scale. Whiskers represent 95% confidence intervals based on standard errors clustered at the district level.

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



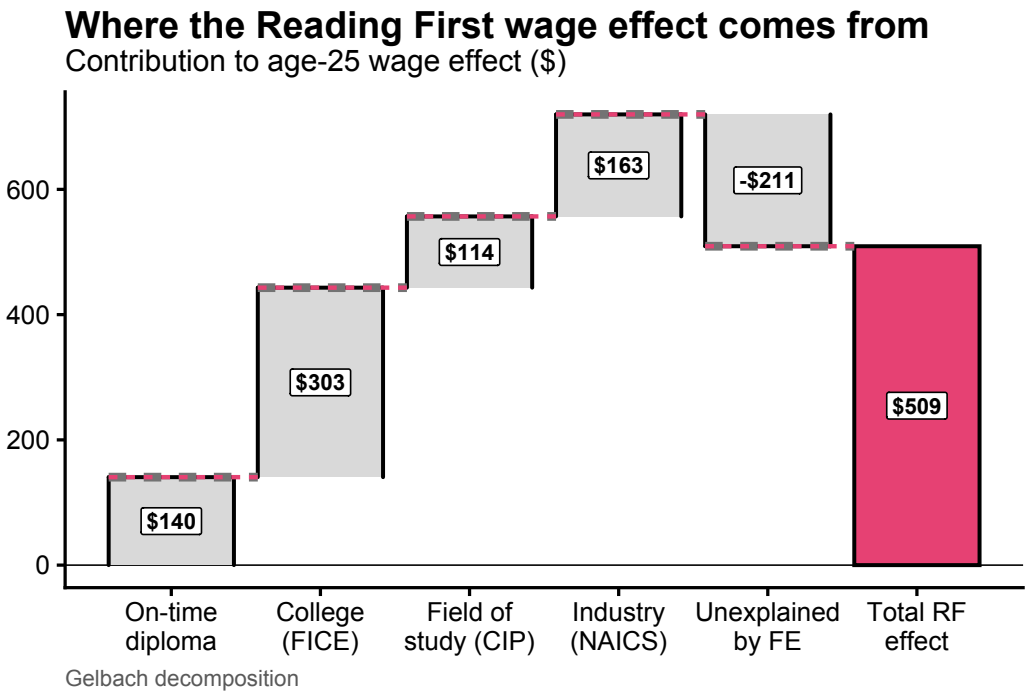
Notes: We consider the distribution of the district-level share of expenditures on supplies & materials and on salary, wages & benefits for districts that received Reading First funding. We split the sample into three bins. The first contains districts below the 2.5th percentile of that distribution, the second contains districts between the 2.5th and 97.5th percentiles, and the third contains districts above the 97.5th percentile. Each point displayed is a difference-in-differences estimate of the effect that Reading First eligibility had on third-grade reading test scores. That effect is estimated separately for each of the three bins across the two expenditure categories. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

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



Notes: Each panel reports estimates of the effect of Reading First eligibility on standardized reading and math scores, high school outcomes, college outcomes, labor force participation, and wages. Reading, math, high school, and college outcomes are estimated by OLS (in standard deviations or as linear probabilities), and labor force participation and wages are estimated by a Poisson model (log points). Reading and math outcomes include all third-grade cohorts. High school, labor force, and wage outcomes use cohorts through 2014. College enrollment outcomes use cohorts observed to age 19 (third-grade cohorts through 2014), and degree outcomes use cohorts observed to age 22 (through 2011). "Ever" counts any enrollment or degree in the data, at any age, while "by 19" and "by 22" count only those that occur by that age. A student is included only if they are observed through age 19 (enrollment) or age 22 (degrees). Within each panel, circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

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



Notes: This figure uses a Gelbach (2016) decomposition to split the total effect of Reading First eligibility on annual wages at age 25 ($509, rightmost bar) into the contributions of intermediate outcomes: on-time diploma receipt ($140), college attended (FICE fixed effects, $303), field of study (CIP fixed effects, $114), industry of employment (NAICS fixed effects, $163), and the component unexplained by these outcomes (−$211).

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

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

Notes: This table reports estimates of the effect of Reading First eligibility on the Reading First revenue that school districts received. The sample is a district–year panel for 1997–2017. In Panels A and B, the outcome is Reading First revenue per K–3 student. Panel A reports the pooled difference-in-differences estimate. Panel B splits the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Panel C reports the first stage for our instrumental variables estimates, a regression of cumulative realized Reading First revenue on cumulative predicted Reading First dollars, both in millions of dollars, along with the first-stage F statistic. This analysis includes district and year fixed effects, and regressions are weighted by K–3 enrollment. Column (1) includes all districts, and column (2) restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

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

Notes: This table reports estimates of the effect of Reading First eligibility on the standardized third-grade reading scores of first-time test takers. The sample is a balanced panel of campuses observed in every year from 1995 to 2017. Panel A reports the pooled difference-in-differences estimate. Panel B splits the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Panel C reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars. The coefficients in Panel C are scaled per $20 million. This analysis includes campus and year fixed effects and the campus's 2002 economically disadvantaged share interacted with year fixed effects, and regressions are weighted by the number of reading test takers. Column (1) includes all districts, and column (2) restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Effect of Reading First eligibility on the third-grade share of correct responses, by question feature

	Pooled		Reading		Math	
	All	BW < 5	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Reading First effect						
1(RF Eligible) × 1(year>2003)	2.308***	1.715***	2.323***	1.665***	2.259***	1.713**
	(0.540)	(0.618)	(0.464)	(0.539)	(0.613)	(0.710)
Panel B. Effect by question feature: 1(RF Eligible) × 1(year>2003) × feature						
Word count (per word)	0.041***	0.016***	0.038***	0.021***	0.039***	0.011
	(0.005)	(0.005)	(0.005)	(0.006)	(0.008)	(0.008)
Share of basic-phonics words (per pct. pt.)	0.029***	0.020***	0.023***	0.009	0.033***	0.035***
	(0.005)	(0.007)	(0.005)	(0.006)	(0.009)	(0.013)
Vocabulary question (= 1)			-0.412***	-0.157		
			(0.158)	(0.159)		
Comprehension question (= 1)			1.307***	0.916***		
			(0.148)	(0.185)		
Answer options are all numbers (= 1)					0.273	0.027
					(0.185)	(0.231)
Observations	3,053,732	1,722,818	1,443,812	814,724	1,609,920	908,094
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell is a separate regression of the percent of third graders at a campus answering a question correctly. Panel A is the difference-in-differences effect of Reading First eligibility. Each row of Panel B adds one feature of the question and reports how the Reading First effect changes with that feature (the triple interaction of eligibility, the post-period, and the feature). The regressions also control for the feature and its two-way interactions with eligibility and the post-period, which are not reported. Rows marked (= 1) are indicators, so their coefficients compare the Reading First effect on questions with and without the question feature. All regressions include campus and year fixed effects and the campus's 2002 economically disadvantaged share interacted with year fixed effects. Each observation is a campus, exam year, and question, weighted by the number of test takers. Word count is counted over the question and its answer options; the basic-phonics share is the share of those words at CIERA decodability levels 1–4; a question's answer options are all numbers when each of the four choices is a bare number, amount of money, time, or fraction (28% of math questions). Vocabulary and comprehension are reading objectives and are reported for reading only; the numeric answer only indicator is reported for math only. The phonics row excludes the few math questions without digitized text (about 10,400 fewer observations in the pooled and math All columns and 5,800 fewer in their BW columns; reading is unaffected). "BW < 5" restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

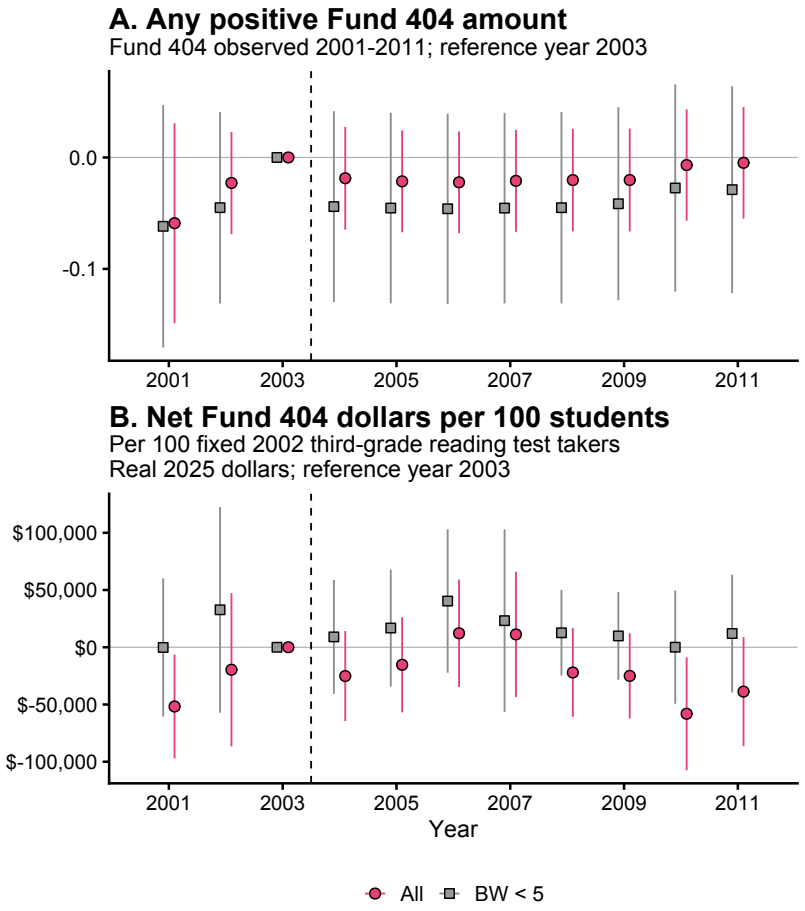
Online Appendix

Decoding Success: Phonics-based Instruction and Student Achievement

Alex Hollingsworth, Alberto Ortega, and Nikolaos Zirogiannis

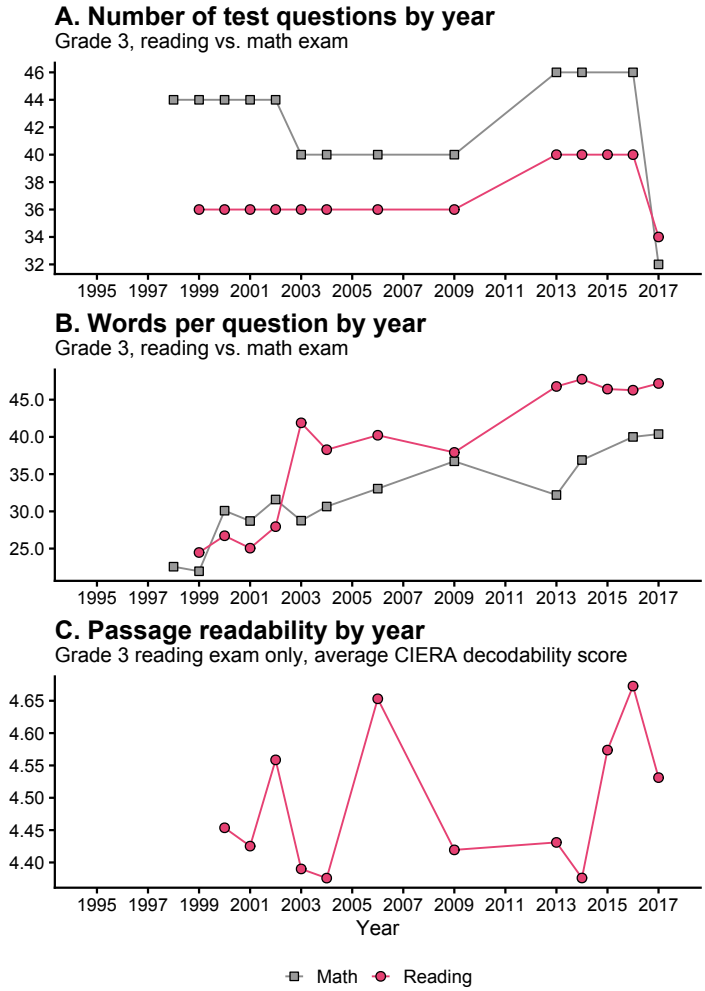
A Appendix Figures

Figure A1: Student Success Initiative funding event studies



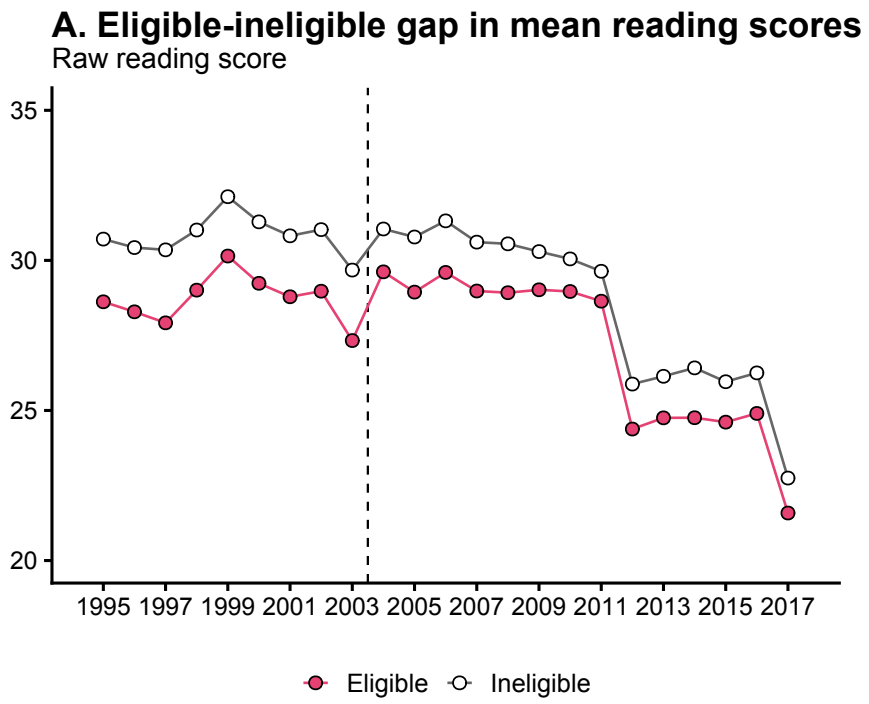
Notes: Event study estimates of the effect of Reading First eligibility on two Student Success Initiative funding outcomes, any positive Fund 404 amount (Panel A) and net Fund 404 dollars per 100 fixed 2002 third-grade reading test takers (Panel B). Fund 404 is the Texas Education Agency's dedicated fund for the initiative. Estimates are relative to 2003, the same reference year as the Reading First event studies. Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A2: Test-question trends for third-grade exams by year



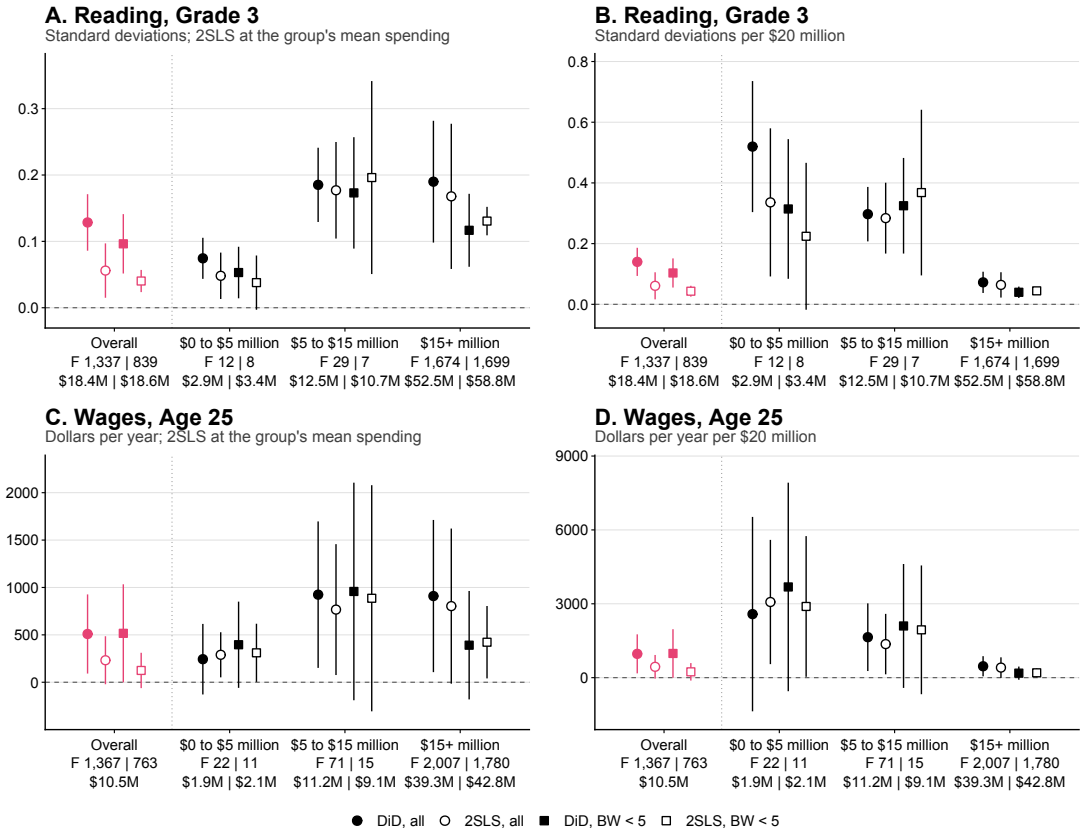
Notes: Panel A plots the number of released third-grade reading and math exam questions available by year. Panel B plots the average number of words per question for reading and math exams. Panel C plots the average decodability score of the words in reading exam questions, using the Center for the Improvement of Early Reading Achievement (CIERA) scale, which runs from 1 (easiest to decode) to 8 (hardest). These questions are digitized from released exams in Texas's TAAS (1995–2002), TAKS (2003–2011), and STAAR (2012 onward) testing programs.

Figure A3: Trend in third-grade reading by Reading First eligibility



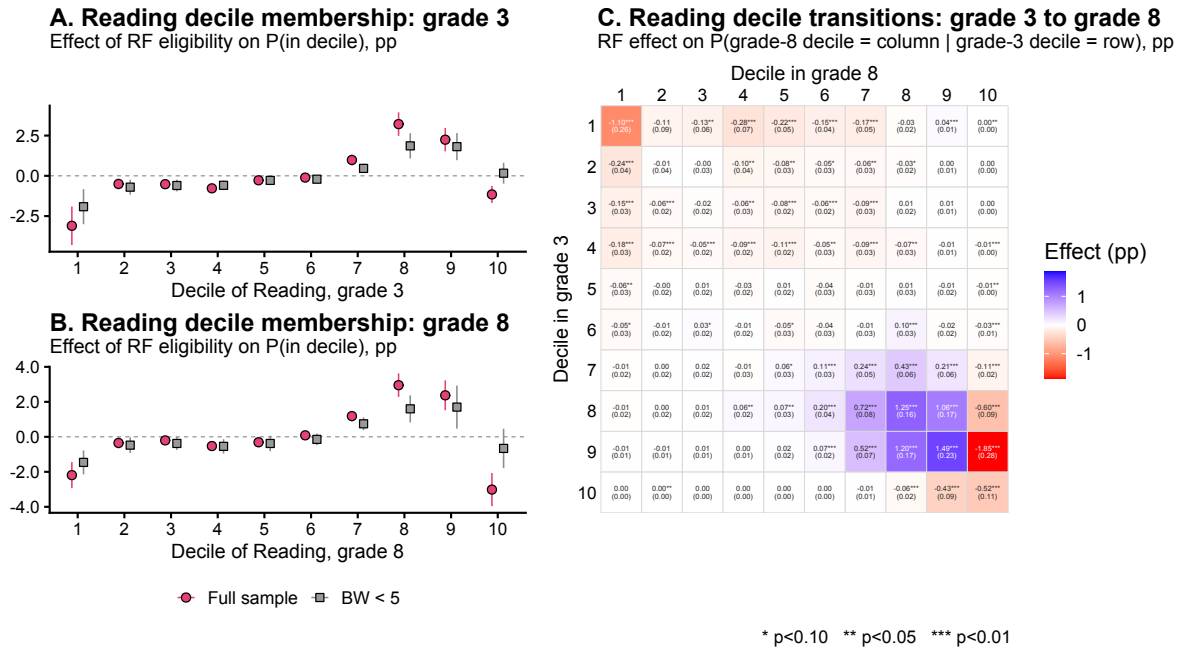
Notes: Mean raw (unstandardized) third-grade reading scores over time for districts eligible and ineligible for Reading First. These are the unadjusted levels behind Panel A of Figure 2.

Figure A4: Treatment effects by three groups of cumulative predicted Reading First dollars, estimated at the group's mean spending and per $20 million



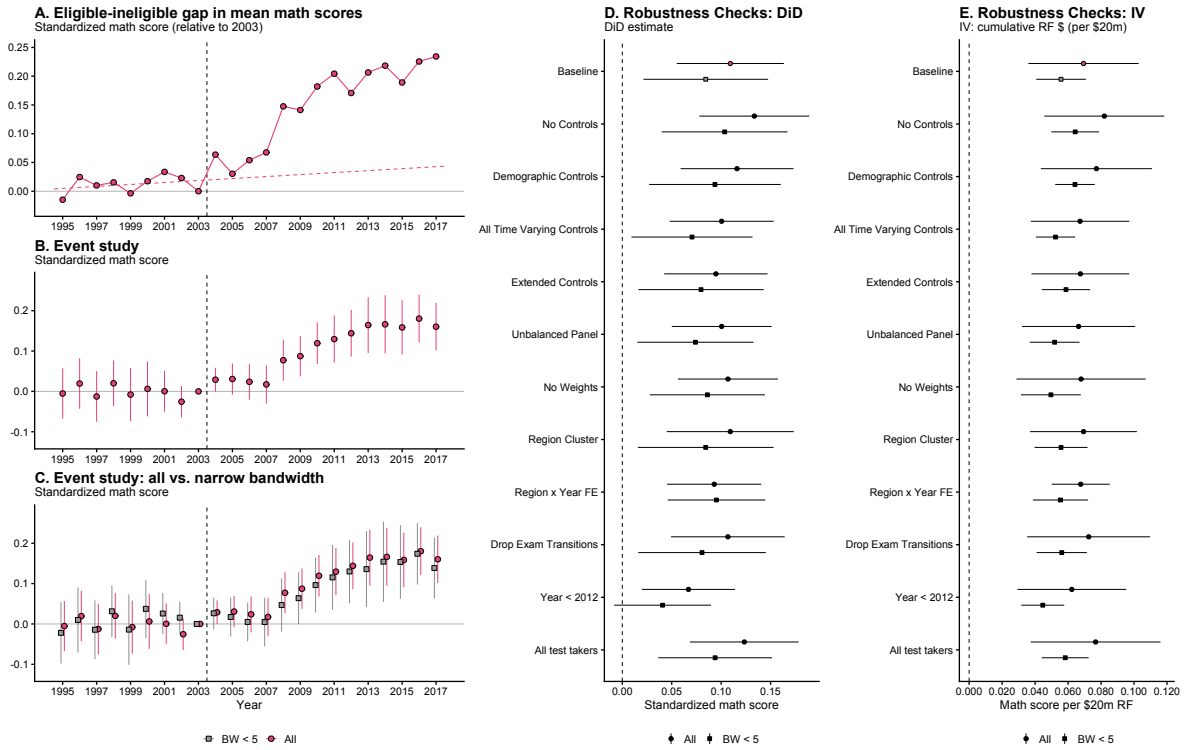
Notes: Each panel reports difference-in-differences (filled markers) and instrumental variables (hollow markers) estimates of the effect of Reading First on the indicated outcome, separately for eligible districts in three groups of cumulative predicted Reading First dollars. The groups are $0 to $5 million (338 districts), $5 to $15 million (16 districts), and $15 million or more (5 districts). Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. The leftmost estimates in each panel, labeled "Overall," pool all eligible districts. Each estimate comes from a separate regression that compares the group's eligible districts after 2003 with all untreated observations, which are ineligible districts in every year and all eligible districts through 2003. In Panels A and C, the instrumental variables estimate is multiplied by the group's mean realized dollars. Below each group, the first-stage F-statistic and the mean realized dollars are reported for the full sample (left of vertical bar) and for the restricted sample (right of vertical bar). Panels B and D instead show every estimate as the effect of $20 million in Reading First funding. Wage panels use third-grade cohorts through 2014, and the sample matches Appendix Figure A17. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A5: Effect of Reading First eligibility on reading decile membership and transitions, grades 3–8



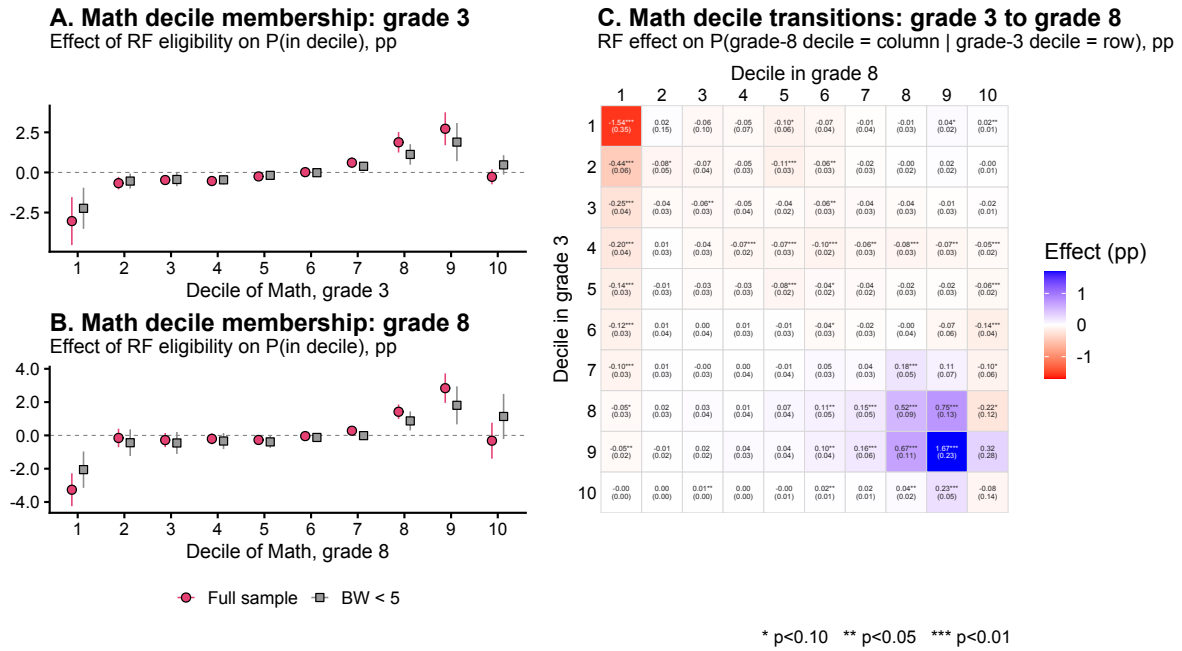
Notes: Panels A and B report difference-in-differences estimates of the effect of Reading First eligibility on third-grade and eighth-grade reading decile membership, in percentage points. Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Panel C reports the effect on the probability of each transition from a third-grade decile (rows) to an eighth-grade decile (columns). Standard errors are clustered at the district level. Bars in Panels A and B represent 95% confidence intervals, and Panel C reports standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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



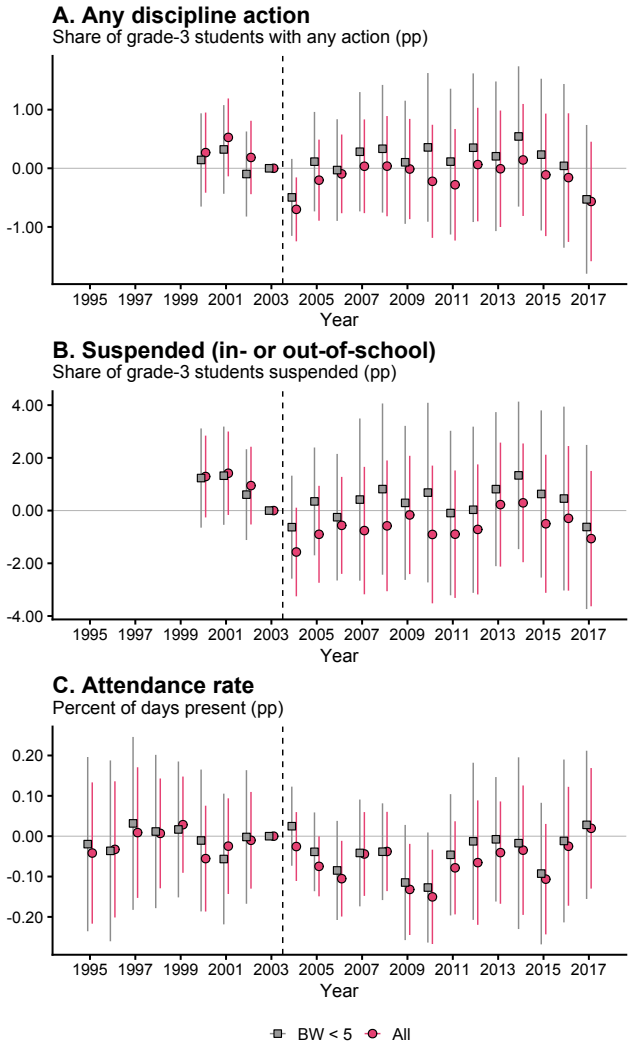
Notes: Panel A plots the raw difference in standardized third-grade math scores between districts eligible and ineligible for Reading First, relative to 2003. Panel B presents event study estimates from equation (3) with 95% confidence intervals. Panel C adds a restricted sample of districts within five percentage points of the 2002 statewide average of 13.6% (square markers). Panel D reports specification checks for the difference-in-differences model, and Panel E reports the corresponding checks for the instrumental variables model, each for the full and restricted samples (see Section 4.2 for details). All models include campus-by-grade and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects. Standard errors clustered at the district level. All samples are restricted to first-time test takers, except for the bottom row of Panels D and E, which uses all test takers. Panel E estimates are per $20 million of cumulative Reading First funding.

Figure A7: Effect of Reading First eligibility on math decile membership and transitions, grades 3–8



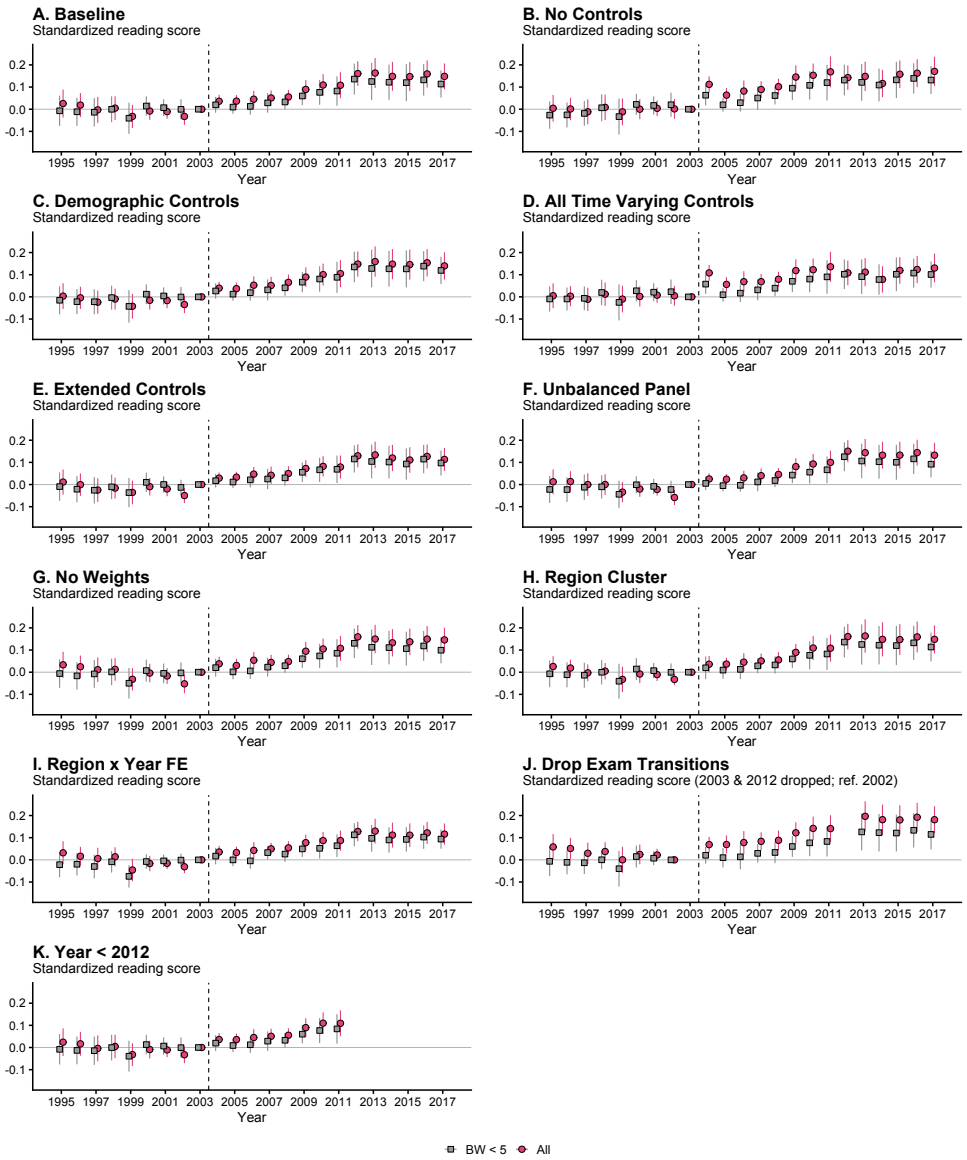
Notes: Panels A and B report difference-in-differences estimates of the effect of Reading First eligibility on third-grade and eighth-grade math decile membership, in percentage points. Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Panel C reports the effect on the probability of each transition from a third-grade decile (rows) to an eighth-grade decile (columns). Standard errors are clustered at the district level. Bars in Panels A and B represent 95% confidence intervals, and Panel C reports standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A8: Event studies for behavioral outcomes



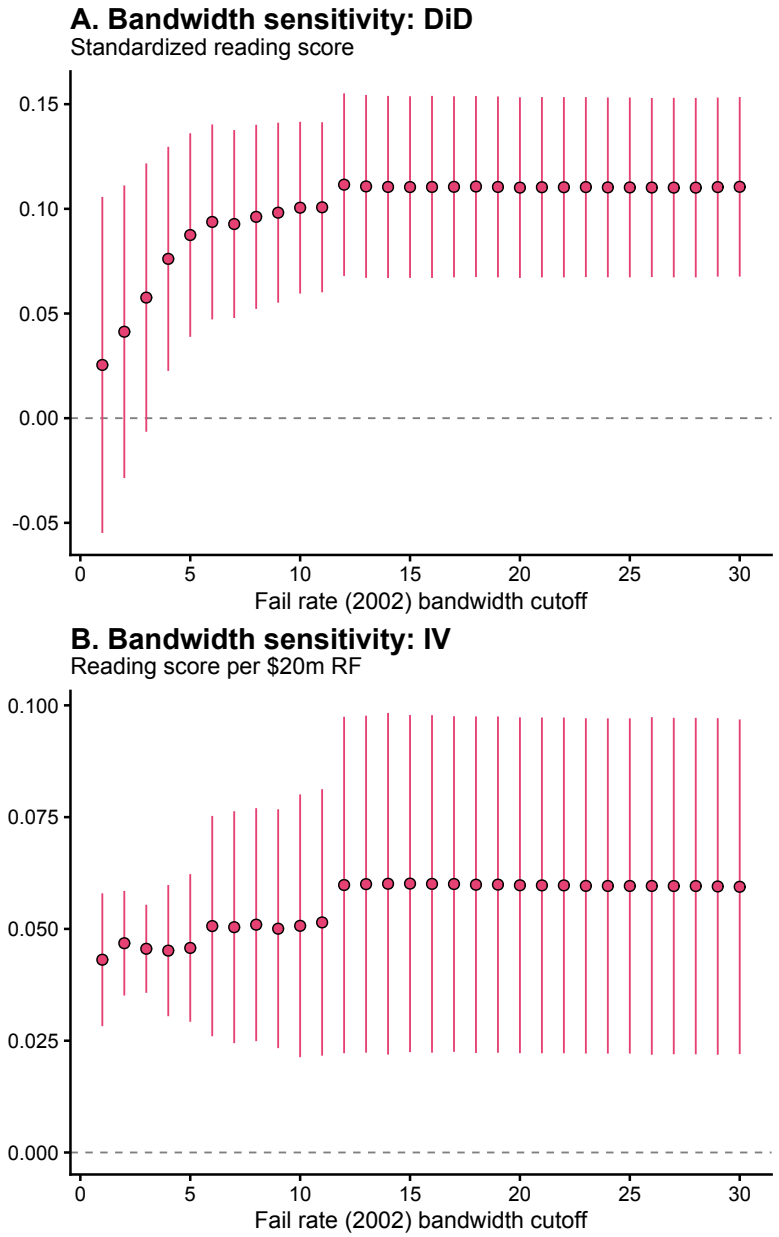
Notes: Event study estimates (relative to 2003) of the effect of Reading First eligibility on three behavioral outcomes, the share of students with any disciplinary action (Panel A), the share of students with in- or out-of-school suspensions (Panel B), and the attendance rate (Panel C). Attendance data are available for the entire sample period, while disciplinary records begin in 2000. Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A9: Event studies under each difference-in-differences robustness specification



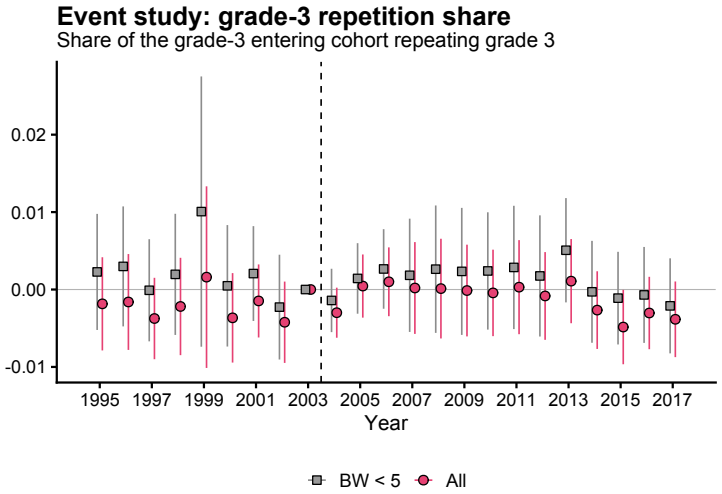
Notes: Each panel reports event study estimates (relative to 2003) of the effect of Reading First eligibility on the standardized third-grade reading scores of first-time test takers under one of the difference-in-differences robustness specifications described in Section 4.2. Panel A is the baseline specification, which includes campus-by-grade and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects. Each subsequent panel is a separate specification denoted in the panel title. Panel J drops the exam transition years 2003 and 2012 and uses 2002 as the reference year. Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A10: Bandwidth sensitivity of the difference-in-differences and instrumental variables estimates



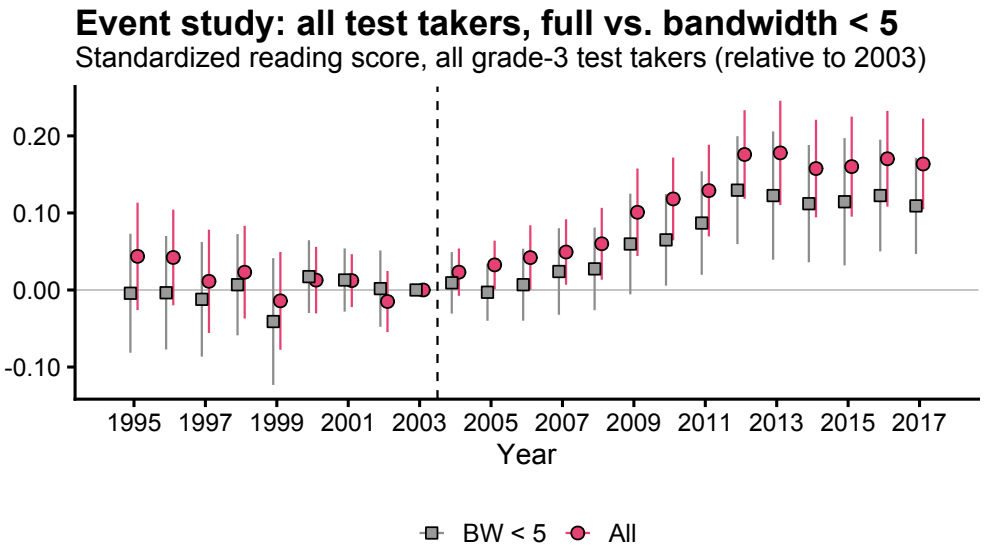
Notes: Each point reports the estimated effect on standardized third-grade reading scores for the sample restricted to districts within the indicated bandwidth (in percentage points) of the 2002 statewide average of 13.6%. Panel A reports the pooled difference-in-differences estimate of the effect of Reading First eligibility. Panel B reports the instrumental variables estimate of the effect per $20 million of cumulative Reading First funding. All models include campus-by-grade and year fixed effects and the campus's fixed 2002 economically disadvantaged share interacted with year fixed effects. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A11: Event study of the third-grade repetition share



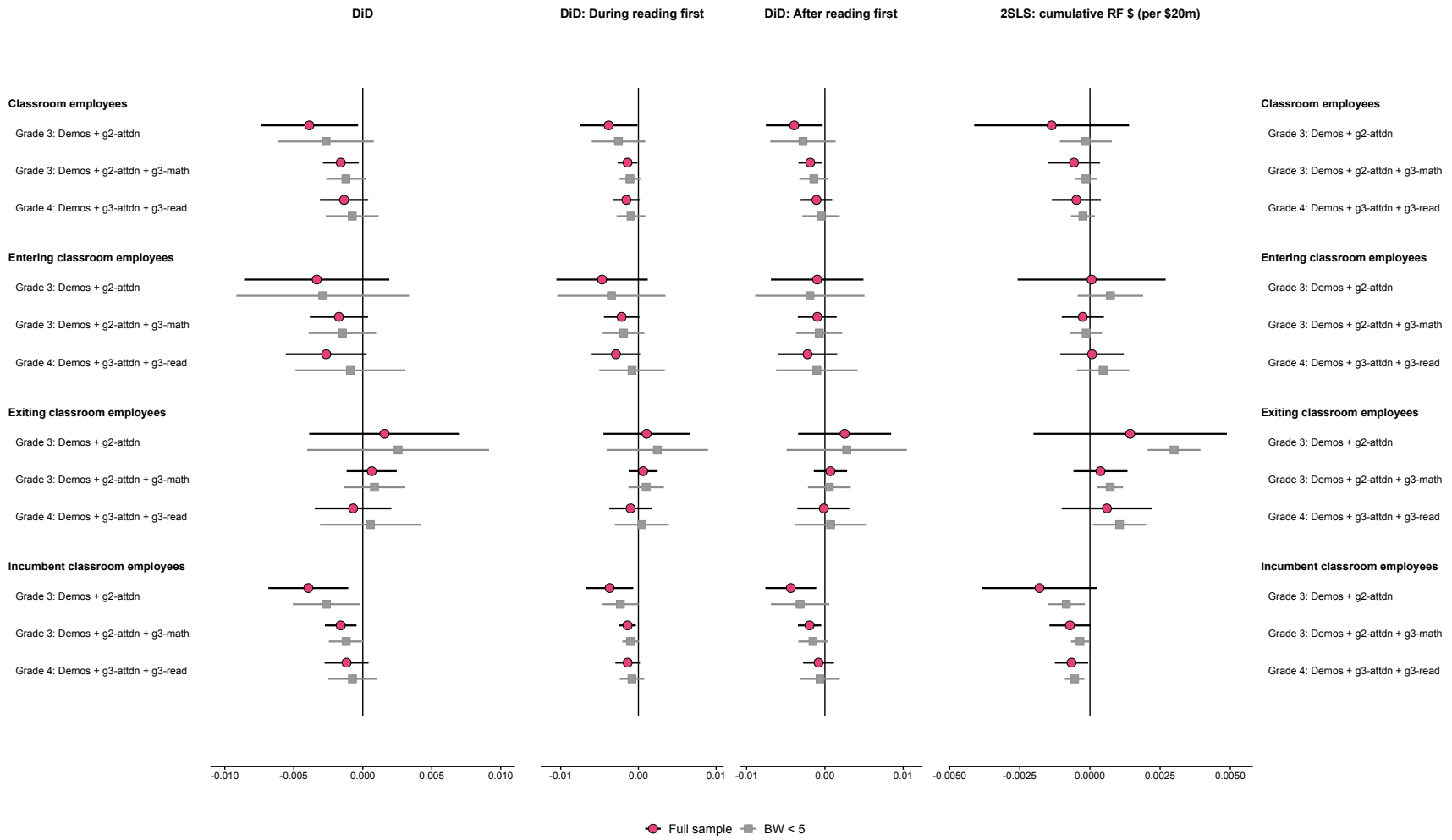
Notes: Event study estimates (relative to 2003) of the effect of Reading First eligibility on the share of the entering third-grade cohort that is repeating third grade, 1995–2017. Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A12: Event study of third-grade reading among all test takers



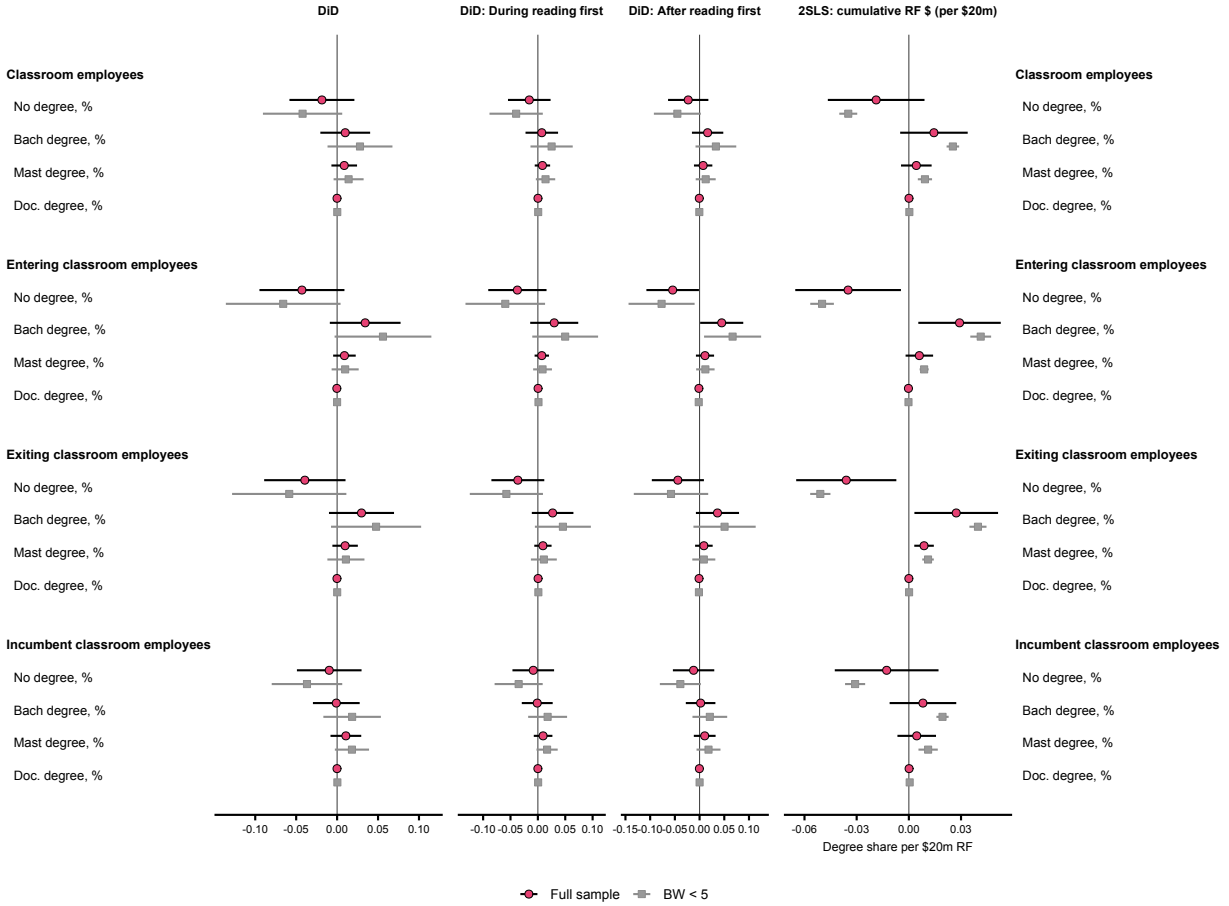
Notes: Event study estimates (relative to 2003) of the effect of Reading First eligibility on standardized third-grade reading scores among all third-grade test takers, including students repeating the grade. Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A13: Changes in average teacher value added by campus



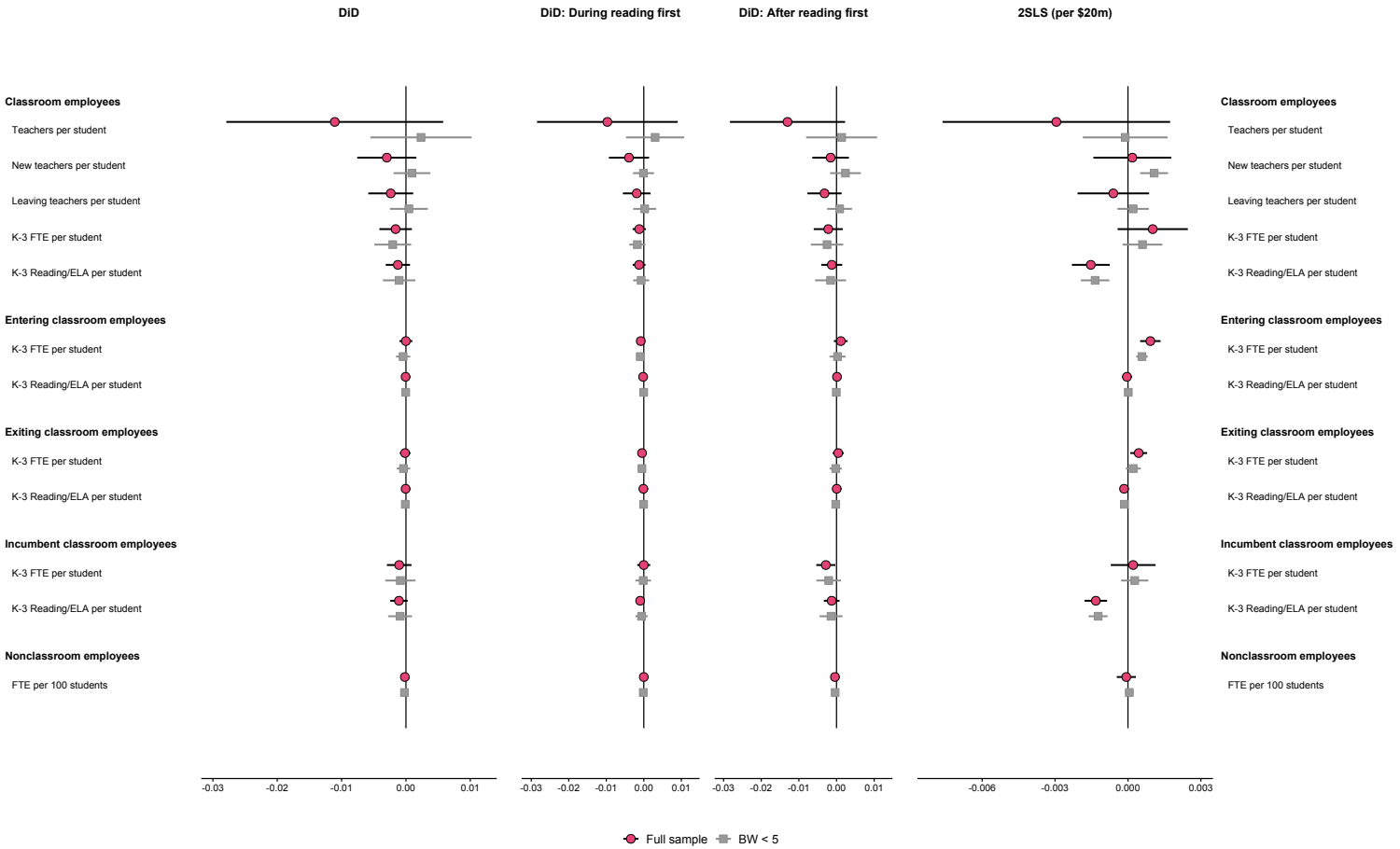
Notes: Each column reports estimates of the effect of Reading First eligibility on campus-level average pre-period employee value added. The first three columns report difference-in-differences estimates, pooled and then separately for the funding years (2004–2011) and the years after funding ended, and the fourth column reports instrumental variables estimates per $20 million of cumulative Reading First funding. Three separate value-added measures are estimated using student-level data for 1995–2002. The first regresses standardized third-grade reading scores on student demographics and second-grade attendance, the second adds the student's third-grade math score, and the third regresses standardized fourth-grade reading scores on fourth-grade demographics, third-grade attendance, and the student's third-grade reading score. Residuals are averaged to campus–year means and shrunk with a standard empirical Bayes procedure (see Section 5.3). Teachers are assigned the value-added score of their 2002 campus and followed over time. Results are shown for all classroom employees, entering teachers, exiting teachers, and incumbents. Within each row, circle markers indicate estimates using the full sample, and the square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A14: Changes in the share of teachers with degrees by campus



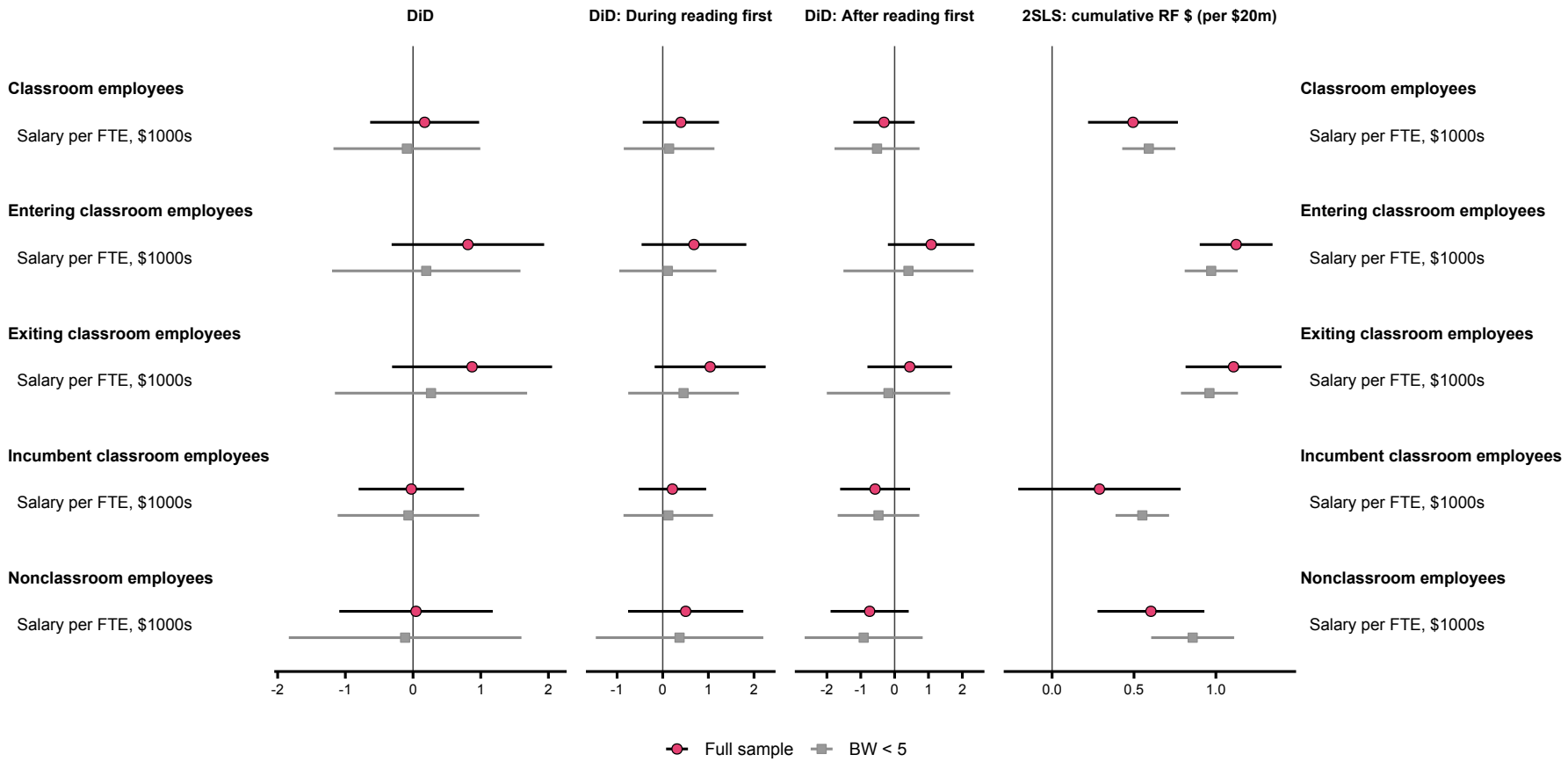
Notes: Each column reports estimates of the effect of Reading First eligibility on the share of classroom employees by degree attainment. The first three columns report difference-in-differences estimates, pooled and then separately for the funding years (2004–2011) and the years after funding ended, and the fourth column reports instrumental variables estimates per $20 million of cumulative Reading First funding. Results are shown for all classroom employees and for entering, exiting, and incumbent employees. Within each row, circle markers indicate estimates using the full sample, and the square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A15: Changes in employees by campus



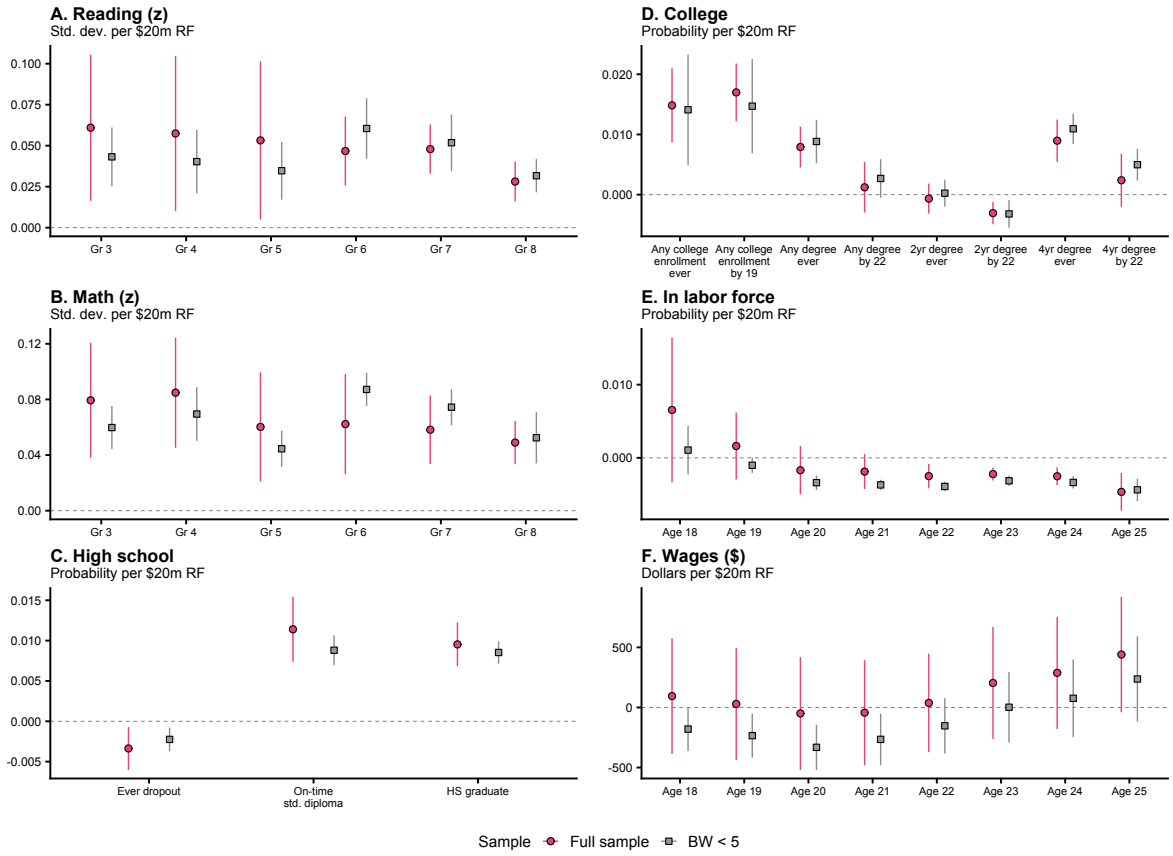
Notes: Each column reports estimates of the effect of Reading First eligibility on various measures of full-time-equivalent (FTE) teachers per student. The first three columns report difference-in-differences estimates, pooled and then separately for the funding years (2004–2011) and the years after funding ended, and the fourth column reports instrumental variables estimates per $20 million of cumulative Reading First funding. Outcomes include the overall teacher-to-student ratio, entering and exiting teachers, K–3 FTE teachers per student, K–3 reading/ELA FTE teachers per student, and FTE nonclassroom employees per 100 students. Within each row, circle markers indicate estimates using the full sample, and the square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A16: Changes in campus salary per FTE employee



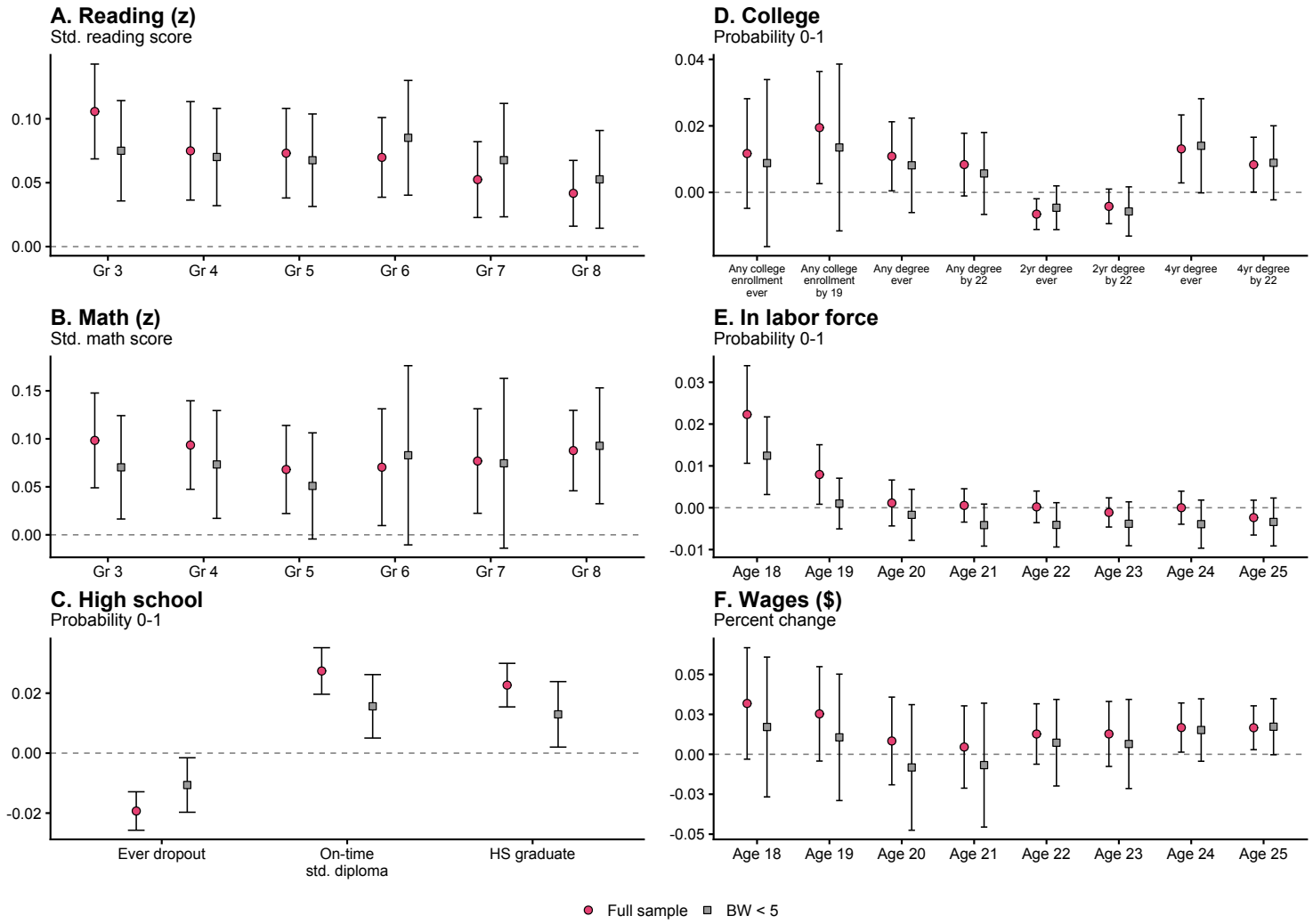
Notes: Each column reports estimates of the effect of Reading First eligibility on salary per FTE employee (in thousands of dollars). The first three columns report difference-in-differences estimates, pooled and then separately for the funding years (2004–2011) and the years after funding ended, and the fourth column reports instrumental variables estimates per $20 million of cumulative Reading First funding. Results are shown separately for all classroom employees and for entering, exiting, incumbent, and nonclassroom employees. Within each row, circle markers indicate estimates using the full sample, and the square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A17: Instrumental variables estimates across the life course, per $20 million of cumulative Reading First funding



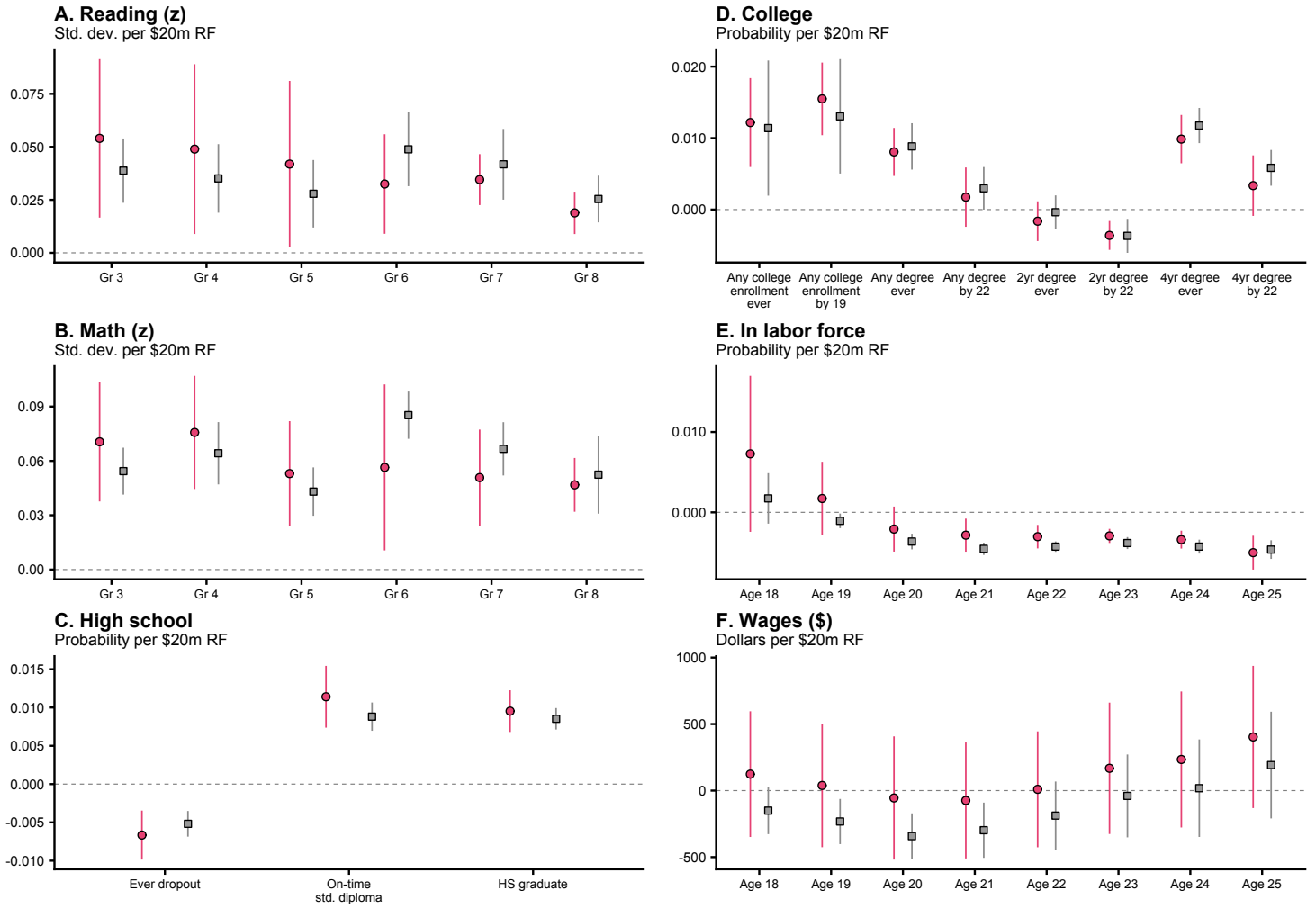
Notes: Each panel reports instrumental variables estimates of the effect of $20 million of cumulative Reading First funding on standardized reading and math scores by grade, high school outcomes, college outcomes, labor force participation, and wages by age, for students observed continuously in Texas public schools from third grade through high school. All estimates come from linear instrumental variables regressions, so effects are in standard deviations, probabilities, or dollars. Reading and math outcomes include all third-grade cohorts, while high school, labor force, and wage outcomes use cohorts through 2014, college enrollment outcomes use cohorts observed to age 19, and degree outcomes use cohorts observed to age 22. Within each panel, circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A18: Long-run effects of Reading First eligibility across the life course, continuously enrolled students



Notes: Each panel reports estimates of the effect of Reading First eligibility on the indicated outcome for students observed continuously in Texas public schools from third grade through high school. Reading, math, high school, and college outcomes are estimated by OLS (standard deviations or linear probabilities), and labor force participation and wages by a Poisson model (log points). Reading and math outcomes include all third-grade cohorts, high school and labor market outcomes use cohorts through 2014, and college outcomes use cohorts observed to age 19 (enrollment) or age 22 (degrees). Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

Figure A19: Instrumental variables estimates across the life course, continuously enrolled students



Notes: Each panel reports instrumental variables estimates of the effect of $20 million of cumulative Reading First funding on the indicated outcome for students observed continuously in Texas public schools from third grade through high school. Effects are in standard deviations, probabilities, or dollars. Reading and math outcomes include all third-grade cohorts, high school and labor market outcomes use cohorts through 2014, and college outcomes use cohorts observed to age 19 (enrollment) or age 22 (degrees). Circle markers indicate estimates using the full sample, and square markers indicate estimates for the sample restricted to districts within five percentage points of the 2002 statewide average of 13.6%. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

B Appendix Tables

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

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

Notes: This table reports average campus and district characteristics before Reading First, over 1995–2002. Column (1) includes all campuses and districts, eligible and ineligible together, and column (2) keeps only districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Panel A describes third graders in the balanced panel of campuses, with means weighted by enrollment and weighted standard deviations in parentheses. Panel B uses each district's 1995–2002 average, with one observation per district, and weights by K–3 enrollment.

Table B2: Effect of Reading First eligibility on Student Success Initiative margins

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

Notes: This table reports estimates of the effect of Reading First eligibility on Student Success Initiative funding, which districts received through Fund 404, the Texas Education Agency's dedicated fund for the initiative. The first row uses an indicator for any positive Fund 404 amount, and the second row uses net Fund 404 dollars per 100 third-grade reading test takers in 2002. Each estimate comes from a separate regression of the outcome on Reading First eligibility interacted with an indicator for 2003 and later, the year the initiative began. The sample is a district–year panel for 2001–2011, and district–years with no Fund 404 record are coded as zero. This analysis includes district and year fixed effects, and regressions are weighted by the district's number of third-grade reading test takers in 2002. Column (1) includes all districts, and column (2) restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: First-stage estimates across robustness specifications

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

Notes: This table reports the first stage for each instrumental variables robustness specification. Each coefficient comes from a regression of cumulative realized Reading First dollars on cumulative predicted Reading First dollars, with the instrument F statistic reported below it. Panel A is the baseline specification, which includes campus and year fixed effects and the campus's 2002 economically disadvantaged share interacted with year fixed effects and is weighted by the number of reading test takers. Each remaining panel changes one feature of the baseline, as named in the panel title and described in section 4.2. Column (1) includes all districts, and column (2) restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district, except in Panel H, where they are clustered by Education Service Center region. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4: Effect of Reading First eligibility on non–Reading First school district revenue

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

Notes: This table reports estimates of the effect of Reading First eligibility on district revenue from sources other than Reading First. The outcome is non–Reading First revenue per student, using total district enrollment in the same year, and the sample is a district–year panel for 1997–2017. Columns (1) and (2) report the pooled difference-in-differences estimate, and columns (3) and (4) split the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). This analysis includes district and year fixed effects, and regressions are weighted by total district enrollment. Odd columns include all districts, and even columns restrict the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B5: Effect of Reading First with control for non–Reading First district revenue (robustness check)

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

Notes: This table reports estimates of the effect of Reading First eligibility on the standardized third-grade reading scores of first-time test takers when we control for other district revenue. Column (1) repeats the baseline specification for 1995–2017. Columns (2)–(4) use 1997–2017, the years with district finance data, with column (3) adding non–Reading First total district revenue as a control and column (4) adding non–Reading First federal revenue. The difference between columns (1) and (2) comes from dropping 1995 and 1996, not from the revenue controls. Revenue controls are measured per K–3 student in thousands of dollars. Panel A reports the pooled difference-in-differences estimate, and Panel B reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars. The coefficients in Panel B are scaled per $20 million. This analysis includes campus and year fixed effects and the campus's 2002 economically disadvantaged share interacted with year fixed effects, and regressions are weighted by the number of reading test takers. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B6: Effect of Reading First eligibility on third-grade reading, all test takers

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

Notes: This table reports estimates of the effect of Reading First eligibility on the standardized third-grade reading scores of all students who took the third-grade reading exam in a given year, including retakes and students repeating third grade. The sample is a balanced panel of campuses observed in every year from 1995 to 2017. Columns (1) and (2) report the pooled difference-in-differences estimate, and columns (3) and (4) split the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Columns (5) and (6) report OLS regressions on cumulative realized Reading First dollars, columns (7) and (8) report the reduced form on cumulative predicted Reading First dollars, and columns (9) and (10) report instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars. The dollar coefficients are scaled per $20 million. This analysis includes campus and year fixed effects and the campus's 2002 economically disadvantaged share interacted with year fixed effects, and regressions are weighted by the number of reading test takers. Within each pair of columns, the first includes all districts, and the second restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B7: Effect of Reading First eligibility on third-grade math

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

Notes: This table reports estimates of the effect of Reading First eligibility on the standardized third-grade math scores of first-time test takers. The sample is a balanced panel of campuses observed in every year from 1995 to 2017. Panel A reports the pooled difference-in-differences estimate. Panel B splits the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Panel C reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars. The coefficients in Panel C are scaled per $20 million. This analysis includes campus and year fixed effects and the campus's 2002 economically disadvantaged share interacted with year fixed effects, and regressions are weighted by the number of math test takers. Column (1) includes all districts, and column (2) restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

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

Notes: This table reports estimates of the effect of Reading First eligibility on the standardized third-grade reading scores of first-time test takers, separately for campuses with above-median and below-median fine particulate matter ($PM_{2.5}$) concentrations. The sample is the balanced panel of campuses observed in every year from 1995 to 2017. We split campuses at the median of their average $PM_{2.5}$ over 2000–2016, the years with satellite $PM_{2.5}$ data (Di et al., 2016), taken across campuses in the reading sample. Panel A reports the pooled difference-in-differences estimate. Panel B splits the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Panel C reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars. The coefficients in Panel C are scaled per $20 million. This analysis includes campus and year fixed effects and the campus's 2002 economically disadvantaged share interacted with year fixed effects, and regressions are weighted by the number of reading test takers. Within each pollution group, the first column includes all districts, and the second restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

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

Notes: This table reports estimates of the effect of Reading First eligibility on how districts spent their Reading First funds. The outcome is Reading First expenditure per K–3 student, and the sample is a district–year panel for 1997–2017. Columns (1) and (2) use all tracked expenditures, and the remaining columns split that total into salary, wages, and benefits (columns (3) and (4)), travel (columns (5) and (6)), professional purchased services (columns (7) and (8)), supplies and materials (columns (9) and (10)), and miscellaneous operating costs (columns (11) and (12)). Panel A reports the pooled difference-in-differences estimate. Panel B splits the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Panel C reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars, scaled per $20 million, along with the first-stage F statistic. This analysis includes district and year fixed effects, and regressions are weighted by K–3 enrollment. Within each outcome, the first column includes all districts, and the second restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B10: Effect of Reading First eligibility on Reading First school district expenditures on supplies and materials

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

Notes: This table reports estimates of the effect of Reading First eligibility on Reading First spending on supplies and materials. The outcome is Reading First expenditure per K–3 student, and the sample is a district–year panel for 1997–2017. Columns (1) and (2) use total supplies and materials, and the remaining columns split that total into books (columns (3) and (4)), general supplies (columns (5) and (6)), and testing materials (columns (7) and (8)). Panel A reports the pooled difference-in-differences estimate. Panel B splits the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Panel C reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars, scaled per $20 million, along with the first-stage F statistic. This analysis includes district and year fixed effects, and regressions are weighted by K–3 enrollment. Within each outcome, the first column includes all districts, and the second restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

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

Notes: This table reports estimates of the effect of Reading First eligibility on Reading First spending on books. The outcome is Reading First expenditure per K–3 student, and the sample is a district–year panel for 1997–2017. Columns (1) and (2) report the effects on total book spending. The remaining columns (3) through (8) split that total into reading materials, library books, and textbooks, respectively. Panel A reports the pooled difference-in-differences estimate. Panel B splits the post-period into the funding years (2004–2011) and the years after funding ended (2012–2017). Panel C reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars, scaled per $20 million, along with the first-stage F statistic. This analysis includes district and year fixed effects, and regressions are weighted by K–3 enrollment. Within each outcome, the first column includes all districts, and the second restricts the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

	High School		Any College Enrollment		College Degree						Age 25	
					Any		2-year		4-year			
	On-Time Diploma	Any Completion	Ever	By 19	Ever	By 22	Ever	By 22	Ever	By 22	Employed	Earnings ($)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. DID (FEOLS)												
1(RF Eligible) × 1(3rd grade Year ≥ 2004)	2.74***	2.27***	1.29	1.89**	0.83*	0.47	-0.52**	-0.40*	0.98**	0.48	-0.24	$509**
	(0.40)	(0.37)	(0.85)	(0.85)	(0.48)	(0.42)	(0.20)	(0.24)	(0.46)	(0.37)	(0.17)	($213)
B. 2SLS: Cumulative RF $ (per $20 million)												
Cumulative RF $ per $20 million (2SLS)	1.14***	0.95***	1.48***	1.70***	0.79***	0.12	-0.07	-0.31***	0.90***	0.24	-0.47***	$441*
	(0.21)	(0.14)	(0.31)	(0.24)	(0.17)	(0.21)	(0.13)	(0.09)	(0.18)	(0.22)	(0.13)	($245)
First-stage F stat	1,368	1,368	1,390	1,390	1,475	1,475	1,475	1,475	1,475	1,475	1,366	1,366
C. Poisson (Earnings Only)												
Poisson coefficient (log points)	–	–	–	–	–	–	–	–	–	–	–	0.019***
												(0.007)
Implied $ effect at mean	–	–	–	–	–	–	–	–	–	–	–	$641
Observations	4,913,997	4,915,214	5,875,509	5,875,509	4,825,091	4,825,091	4,825,091	4,825,091	4,825,091	4,825,091	3,214,226	3,214,546
Last third-grade cohort	2014	2014	2014	2014	2011	2011	2011	2011	2011	2011	2014	2014
Pre-RF Cohort Mean	90.7	92.9	62.0	56.8	35.9	25.4	11.6	6.9	29.9	20.1	78.7	$33,713
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Third-Grade-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports estimates of the effect of Reading First eligibility on high school, college, and age-25 labor market outcomes. The unit of observation is the student, with students who repeated third grade counted once at their first attempt, and the last third-grade cohort in each column is listed in the table. Columns (1) and (2) report on-time standard high school diplomas and any high school completion for cohorts through 2014. Columns (3) and (4) report college enrollment at any age and by age 19, for students observed through age 19 (cohorts through 2014). Columns (5)–(10) report any degree, a 2-year degree, and a 4-year degree, each at any age and by age 22, for students observed through age 22 (cohorts through 2011). The two columns for each outcome use the same students. Columns (11) and (12) report employment and earnings at age 25 for students whose age-25 year falls in the Texas wage records and who appear in those records at least once between ages 18 and 25. Students with no age-25 wage record are counted as not employed with zero earnings. Coefficients in columns (1)–(11) are in percentage points, and column (12) is in dollars. Panel A reports the pooled difference-in-differences estimate. Panel B reports instrumental variables estimates that use cumulative predicted Reading First dollars as an instrument for cumulative realized Reading First dollars (scaled per $20 million). Panel C reports a Poisson version of the earnings regression, with the implied dollar effect evaluated at the pre–Reading First mean. This analysis includes campus and third-grade-year fixed effects. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B13: Age-25 wages and expected-wage channels

	Actual	Expected wage at 25, given:			Complex
	wage at 25	College (FICE)	Field (CIP)	Industry (NAICS)	prediction
	(1)	(2)	(3)	(4)	(5)
A. Effect of RF on actual and expected wages at 25					
Reading First (post)	509.2**	625.3***	500.8***	230.5***	633.9***
	(213.0)	(122.8)	(90.1)	(89.3)	(145.8)
B. Actual wage at 25, controlling for each expected-wage channel					
Reading First (post)	509.2**	-77.8	29.4	282.5	-43.8
	(213.0)	(205.7)	(189.8)	(175.1)	(160.7)
Expected-wage control	–	0.939***	0.958***	0.983***	0.872***
		(0.007)	(0.008)	(0.010)	(0.006)
Observations	3,214,546	3,214,483	3,214,544	3,214,546	3,214,546
Campus FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports estimates of the effect of Reading First eligibility on actual and expected wages at age 25. The sample includes first-time third-grade test takers matched to Texas unemployment insurance wage records at age 25. The actual wage in column (1) is set to zero for students who appear in the Texas wage records at least once between ages 18 and 25 but have no age-25 wage record. Columns (2)–(4) use the expected wage given the student's college (FICE), field of study (CIP), or industry at age 25 (NAICS), measured as the mean age-25 wage of untreated students in the same category, with separate categories for students who did not attend college or were not employed. Column (5) uses a predicted wage from a Poisson fixed-effects model that combines all three channels. Panel A reports the effect of eligibility on each wage measure. In Panel B, the outcome is the actual age-25 wage in every column, and each column adds its expected-wage measure as a control. Column (1) of Panel B has no expected-wage control and repeats the unadjusted estimate. This analysis includes campus and third-grade-year fixed effects. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C Data and Estimation Details

C.1 Tax Elections

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

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

We can also identify M&O tax rate increases by assessing whether the district's M&O rate rises by at least \$0.11 net of the statewide change in that year. We obtain the \$0.11 threshold by regressing the year-over-year change in a district's M&O rate on a passage indicator and year fixed effects, over the years the Abott et al. (2020) records cover. When applying this threshold back to the districts with observed elections, we can identify the correct passage year for 91% of districts (94% within a year). [25]

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

C.1.1 Finance outcomes

Revenue and expenditure data come from the Public Education Information Management System financial records described in Section 2.3, deflated to 2025 dollars with the CPI-U. PEIMS contains object codes that identify its source and fund and function codes describing where the money is held and how it is to be used. We classify revenue as local, state, or federal using the object code independent of fund or function. [26] Salary, wages and benefits, travel, professional and purchased services, supplies and materials, and miscellaneous operating follow the same object code groupings used for Reading First expenditure. General supplies, testing materials, and books are components of supplies and materials, and books is the sum of library books, textbooks, and reading materials.

The results in Table C1 are estimated at the district–year level. Each expenditure record in PEIMS carries an organization code that assigns it either to a campus or to a district-level unit such as central administration. Panel A sums the revenue records assigned to the district as a

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

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

whole. Panel B sums the expenditure records assigned to campuses, so all tracked expenditure is the district's total campus-assigned spending in the five listed categories rather than total district spending. Note that we cannot observe *where* campus expenditures were allocated, only *that* expenditures were allocated to campuses. Each outcome is divided by the district's total enrollment in that year, from the TEA enrollment files, and the same count serves as the regression weight.

C.1.2 Estimation

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

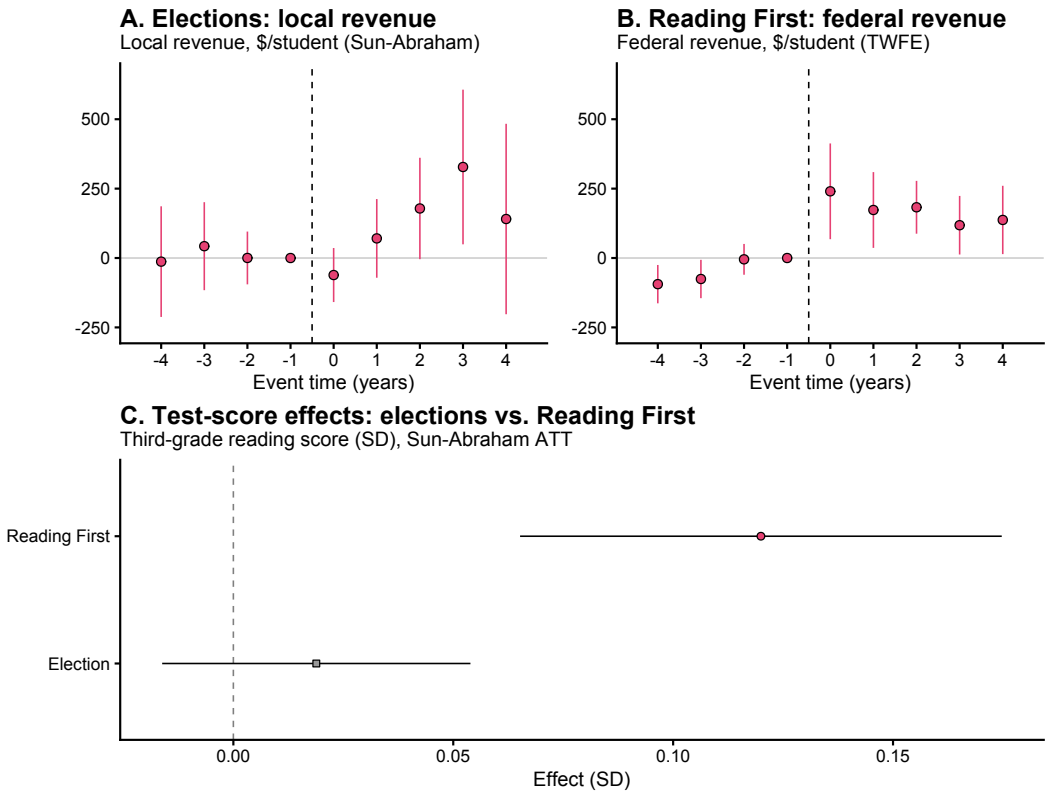
Panel C compares the two treatments on third-grade reading. Both estimates come from the same campus–year panel for 2000–2019 with campus and year fixed effects, weighted by the number of tested students and clustered by district. Passage of a tax increase changes reading scores by 0.019 standard deviations,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

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

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

Notes: This table reports estimates of the effect of passing a maintenance and operations tax election on district revenue and spending per enrolled student. The sample is a district–year panel for 2000–2019, and each estimate is the overall average treatment effect on the treated from the estimator of Sun and Abraham (2021). A district is treated starting in the year it first passes a tax election, and districts that have not yet passed or never pass an election serve as controls. Passage years for 2000–2015 come from the election records of Abott et al. (2020). For 2016–2019, we date passage to the first year a district's maintenance and operations tax rate crosses the $1.04 rollback ceiling or rises by at least $0.11 more than the statewide change, a rule that recovers the recorded year for 91% of districts in the Abott et al. (2020) records. General supplies, testing materials, and books are components of supplies and materials, and books refers to the sum of library books, textbooks, and reading materials. This analysis includes district and year fixed effects, and regressions are weighted by district enrollment. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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



Notes: Panel A reports event study estimates of the effect of passing a school district maintenance and operations tax election on local revenue per enrolled student, using the Sun and Abraham (2021) estimator. Panel B reports event study estimates of the effect of Reading First eligibility on federal revenue per enrolled student, using two-way fixed effects. Both panels use 2000–2019, with event time measured relative to the year before treatment. Panel C compares the implied average treatment effects on standardized third-grade reading scores for tax elections and Reading First. Bars represent 95% confidence intervals based on standard errors clustered at the district level.

C.2 Reading Intensity of Occupations

C.2.1 Occupation scores

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

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

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

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

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

C.2.3 Sample and estimation

The sample here is composed of the same cohorts as those in Table B12, students in third grade through 2014 for high school; those in third grade through 2011 for the college outcomes; and those in third grade through 2009 for the labor market outcomes, the last cohort for whom we observe age-25 outcomes. We estimate the DID specification from Section 3.1 at the student level in Table C2, using campus and third-grade-year fixed effects and standard errors clustered by district. Observation counts differ across outcomes because a student only enters a column if

the relevant choice has a score. In even-numbered columns, the sample is restricted to districts within five percentage points of the eligibility threshold. None of the point estimates in Table C2 are statistically significant, indicating that Reading First did not affect the reading intensity of the choice of CIP, college, or employment industry for treated students.

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

	Field of Study (CIP)		College Attended (FICE)		Industry of Employment (NAICS)	
	All	BW < 5	All	BW < 5	All	BW < 5
	(1)	(2)	(3)	(4)	(5)	(6)
1(RF Eligible) × 1(Year ≥ 2004)	-0.005	-0.005	-0.000	-0.001	-0.000	-0.003
	(0.003)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
Observations	2,590,581	1,388,547	3,004,758	1,611,159	4,740,205	2,604,332
Campus FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports estimates of the effect of Reading First eligibility on the reading intensity of students' field of study, college, and industry of employment. Each outcome is an O*NET reading-importance score in standard deviations across occupations, constructed as described in Appendix C.2. Columns (1) and (2) use the student's field of study at first college enrollment, columns (3) and (4) use the college first attended, and columns (5) and (6) use the student's industry of main employment. The sample includes third-grade cohorts through 2014, and each estimate comes from a student-level difference-in-differences regression with campus and third-grade-year fixed effects. Odd columns include all districts, and even columns restrict the analysis to districts whose 2002 third-grade reading fail rate is within five percentage points of the 2002 statewide average of 13.6%. Robust standard errors are clustered by district. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C.3 Construction of the Marginal Value of Public Funds

To estimate the MVPF for Reading First, we follow Hendren and Sprung-Keyser (2020) throughout.[27] The MVPF is the ratio of beneficiaries' willingness to pay to the net cost borne by the government,

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

where E is the present value of the lifetime earnings gain generated per dollar of program spending and τ is the marginal tax-and-transfer rate. Given that the MVPF is a ratio, it is invariant to scale, so every quantity is normalized to one dollar of government spending. The denominator is a *net* cost, where the government pays out a dollar and recovers τE of this dollar via taxes on the higher future earnings. When τE reaches one, the net cost is zero and implies that the policy has paid for itself.

C.3.1 Inputs

The calculation requires five quantities. First is the earnings effect, which we obtain from our DID estimate of \$509 on earnings at age 25, reported in Table B12, against a mean of \$33,713 among students in untreated cells, or a proportional effect of 1.51%. Both figures are in 2025 dollars.

Second, we calculate program spending per exposed student, which we use to express our earnings estimate per dollar spent. Table 1 reports that eligible districts received \$132.27 per K–3 student per year while program funds were flowing. A student's exposure can run up to

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

four years from K–3. The cohorts contributing to the age-25 earnings estimate were in third grade between 2004 and 2009, and the earliest cohort had already passed through part of K–3 before funds arrived. Measured directly in the data, these cohorts overlapped with the 2004–2011 funding window for an average of 2.72 years. To be conservative, we round the funding window up to three years. Cumulative spending is therefore approximately $397 per exposed student, in the same 2025 dollars as the earnings estimate.

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

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

C.3.2 Lifetime present value

We now extend the single earnings effect from age 25 through age 65. Given that the wage premium effect of the Reading First treatment at age 25 is 1.51%, dividing by the cumulative per student spending of $397 yields an effect of 0.0038%, attributable to a single dollar of spending. We assume that each dollar of program spending was equally productive over time. We then take the mean earnings by age of the 2015 ACS, adjust it to 2025 dollars and to our cohorts' calendar years using 0.5% annual real wage growth, and scale the profile so that it matches our own sample's mean at age 25.

When applying the per dollar earnings effect of 0.0038% to counterfactual mean earnings of $33,713, we obtain $1.28 resulting from a dollar of Reading First spending by age 25. For all ages, we apply the same percentage. Discounting the gains at each age back to age 25 at 3% gives $54.77 per dollar of spending. The money itself was spent when students were about 7, so we discount another eighteen years to compare the costs and benefits at the same date. This leaves a present value of $32.17 in lifetime earnings per dollar of Reading First spending.

The government recovered part of the $32.17 in lifetime earnings per dollar through taxes and reduced transfers. If we apply an 18.9% tax-and-transfer rate, it yields $6.08 of additional revenue per dollar spent on Reading First. Therefore, the net cost is one dollar less $6.08, or −$5.08. Willingness to pay is the earnings gain net of those taxes, $32.17 less $6.08, or $26.09. Because the revenue recovered exceeds the dollar spent, the net cost is negative, and the MVPF is unbounded. Reading First paid for itself.

Table C3 reports how far into adulthood the earnings gains must have persisted for the program to have paid for itself. The first row is most conservative, where we consider only the age 25 earnings. The earnings gain is $1.28 per dollar of spending, discounted to $0.75 at the time of exposure. At this age, the government had recovered fourteen cents of each dollar, for an MVPF of 0.71. Each subsequent row adds one more year of the projected gain; net cost is zero once cumulative recovered revenue equals the dollar spent, which occurred by age 31.

We note that this age 31 threshold is conservative. We hold the proportional earnings effect constant over the life cycle, which existing literature suggests is unlikely. For instance, Mountjoy (2026) estimates the returns to Texas public universities year by year over exactly this range and finds the proportional effect more than doubling between ages 25 and 30. Substituting this trajectory for ages 25–30 and holding the proportional gain constant thereafter brings the "pays for itself" threshold forward to age 29 (columns (6) and (7) of Table C3). Thus, our

constant-effect assumption likely understates the returns to Reading First.

Table C3: Marginal value of public funds by earnings projection horizon

	Constant share of the age-earnings profile					Mountjoy (2026) trajectory	
Earnings projected through age	PV of earnings gain per $1 (1)	Additional tax revenue per $1 (2)	Net cost to government per $1 (3)	Beneficiary WTP per $1 (4)	MVPF (5)	Net cost to government per $1 (6)	MVPF (7)
25 (no projection)	$0.75	$0.14	$0.86	$0.61	0.71	$0.86	0.71
26	$1.54	$0.29	$0.71	$1.25	1.77	$0.66	2.16
27	$2.39	$0.45	$0.55	$1.93	3.52	$0.48	4.71
28	$3.26	$0.62	$0.38	$2.64	6.87	$0.22	14.83
29	$4.16	$0.79	$0.21	$3.38	15.83	**–$0.09**	∞
30	$5.09	$0.96	$0.04	$4.12	105.98	–$0.46	∞
31	**$6.02**	**$1.14**	**–$0.14**	**$4.88**	∞	–$0.83	∞
35	$9.89	$1.87	–$0.87	$8.02	∞	–$2.37	∞
45	$19.67	$3.72	–$2.72	$15.95	∞	–$6.25	∞
55	$27.55	$5.21	–$4.21	$22.34	∞	–$9.38	∞
65 (baseline)	$32.17	$6.08	–$5.08	$26.09	∞	–$11.22	∞

Notes: This table reports the marginal value of public funds (MVPF) for Reading First when the earnings gain is projected through different ages. All columns are per $1 of Reading First spending, following Hendren and Sprung-Keyser (2020). Each row assumes the earnings effect lasts through the stated age and is zero thereafter. Column (1) is the present value of the earnings gain, discounted at 3% to age 7. Column (2) is the added tax revenue at a marginal tax-and-transfer rate of 18.9%. Column (3) is the net cost to the government, or one dollar minus column (2). Column (4) is beneficiaries' willingness to pay, or column (1) minus column (2). Column (5) is the MVPF, or column (4) divided by column (3). Columns (1)–(5) hold the earnings gain at a constant share of the age-earnings profile. Columns (6) and (7) repeat the net cost and the MVPF when, from age 25 to 30, the gain instead grows at the year-over-year growth of the dollar earnings return to Texas public universities in Mountjoy (2026) (Online Appendix Table A.1, years 7–12 after high school graduation), after which the proportional gain is held constant. Bold marks the first horizon at which the program has paid for itself under each trajectory. Appendix C.3 describes the construction and sensitivity checks.

Our estimates are also not sensitive to some of the choices made in the MVPF calculation. The program still pays for itself at discount rates of 1% to 8%, with the threshold moving from age 29 to 45 across that range. We also find similar results for marginal tax rates from 10% to 30%. The same is true for every plausible value of cumulative program spending, including an upper bound that assigns every cohort a full four years of exposure. The MVPF becomes finite only at a discount rate of 9% or above or at cumulative spending more than six times our central estimate.