Check for updates

Research Article

Educational Evaluation and Policy Analysis
Month 202X, Vol. XX, No. X, pp. 1–32
DOI: 10.3102/01623737241303142
Article reuse guidelines: sagepub.com/journals-permissions
© 2024 AERA. https://journals.sagepub.com/home/epa

Effects of Large-Scale Early Math Interventions on Student Outcomes: Evidence From Kentucky's Math Achievement Fund

Zeyu Xu
American Institutes for Research

Umut Özek
RAND

Jesse Levin
Dong Hoon Lee
American Institutes for Research

Addressing the educational needs of students in math early on is critical given that early gaps in math skills widen further over the course of schooling. This study examines the effectiveness and costs of Kentucky's Math Achievement Fund—a unique state-level program that combines targeted interventions, peer-coaching, and close collaboration among teachers to improve math achievement in grades K–3. The program is found to improve not only math achievement, but also reading test scores and non-test outcomes including student attendance and disciplinary incidents. The benefits exist across students from various socioeconomic backgrounds, and they are slightly higher for racial minorities.

Keywords: *early childhood, educational policy, evaluation, mathematics education, quasi-experimental analysis, secondary data analysis*

Introduction

A LONG line of research demonstrates the economic value of mathematics skills to both individuals and the economy as a whole (Behrman et al., 2008; Glewwe, 1996; Hershbein et al., 2014). Based on a review of evidence from the United States (Altonji & Pierret, 2001; Lazear, 2003; Murnane et al., 2000), Hanushek and Woessmann (2008) conclude that one standard deviation increase in mathematics performance at the end of high school translates into 12% higher annual earnings.

Adolescent mathematics performance, in turn, is strongly predicted by students' early grade mathematics skills (Austin et al., 2023; Goldhaber et al, 2021; Watts et al., 2014). Although the association between children's early skills and their later development is present in many domains, such association is twice as strong in mathematics as in reading after controlling for preschool entry cognitive and socioemotional skills and family background (Duncan et al., 2007). More importantly, children who exhibit early delays in mathematical development also demonstrate lower growth rates in mathematics skills later (Baumert et al., 2012; Morgan et al., 2011). Consequently, early gaps in mathematics skills widen further over the course of schooling (Geary et al., 2013; Loeb & Bassok, 2007; Magnuson & Duncan, 2006). How to diagnose and correct mathematics deficiencies early,

therefore, is a challenge that our K–12 schools must confront *systematically*. That said, as pointed out in a National Academies report, attention to early mathematics learning does not match its importance, and reading has been the focus of early childhood education (Cross et al., 2009).[1]

In this study, we examine the effects of Kentucky's Math Achievement Fund (MAF)—a prime example of a state's effort to intervene early to improve student mathematics achievement—on student outcomes using a difference-in-differences (DiD) design. To the best of our knowledge, MAF was the only fully implemented statewide K–3 mathematics intervention program at the time of this study. Kentucky's early grade mathematics performance mirrors the national trend, with roughly half of third-grade students proficient in mathematics in 2017. To help students in grades K–3 who struggle to meet grade-level expectations, MAF adopts a Response to Intervention (RtI) approach that consists of a universal screener, tiered interventions with increasing intensity, and regular progress monitoring. It also requires professional development for both the intervention teacher and classroom teachers in each grantee school. The goal of MAF is to identify mathematics deficiencies or the risks of mathematics difficulties early and provide remedial or preventative interventions before it becomes too late.

Our estimates suggest no statistically significant difference in mathematics achievement trends between MAF and non-MAF schools (i.e., schools that have never implemented MAF) in pretreatment periods, yet the trends started to diverge gradually after MAF designation. For example, we find no effect of MAF designation on mathematics scores at the end of the first year of implementation, but student mathematics test scores in MAF schools rose by 0.05 standard deviation after 2 years relative to non-MAF schools, 0.08 standard deviation after 3 years, and 0.09 standard deviation after 4 years. Evidence also suggests positive spillover effects of the program on reading achievement, with students in MAF schools outperforming non-MAF school students by 0.06 standard deviation after 4 years. Additionally, MAF does not have any unintended negative association with student behaviors that may result from labeling young children as not on track to meet grade expectations in mathematics. On the contrary, relative to non-MAF schools, disciplinary incidents and absences were reduced significantly in MAF schools relative to non-MAF schools. Using detailed information obtained from administrative records and surveys administered to school MAF teams and coordinators, we find that the estimated 4-year per-pupil cost of the MAF intervention is $750 in 2018–2019 dollars, which corresponds to an increase in math test scores by 0.09 standard deviation. As we discuss at the end of this study, closing gaps in mathematics skills early has important economic and equity implications. But even without considering those long-term benefits, these findings suggest that state-led early mathematics intervention programs like MAF could provide a cost-effective blueprint to raise mathematics achievement in early grades.

Policy Background

As provisioned in Kentucky statute KRS 158.844, MAF is a school-level program with two objectives: to provide teacher professional development and to address the needs of students in primary grades (Grades K–3) who are struggling to meet the grade-level expectations for mathematics. The intervention is supposed to supplement, not replace, regular classroom instruction. In addition, MAF was also designed to help schools build capacity by overhauling mathematics instructional strategy and practice and strengthening collaboration among teachers. Through renewable, 2-year local grants to approximately 90 schools per year, MAF awarded $50,000 per school-year to support the hiring of one full-time mathematics intervention teacher (MIT) and training for the MIT and two classroom mathematics teachers (known as the Plus 2 teachers). The anticipated cost to operate an MAF program typically exceeded $50,000, and grantee schools were required to provide additional funds (often from Title I funding) to cover the difference. MAF requires a common structure of intervention with a set of prescribed components, but grantee schools also have discretion over how specific components are implemented, as described below.

Student Selection

All K–3 students in MAF schools are required to take a screener test to determine who are struggling to meet grade-level or benchmark expectations for mathematics and thus need supplemental support. In the 2018–2019 school year, MAF schools in Kentucky primarily used three screeners: Measures of Academic Progress (MAP) in 55 schools, iReady in 18 schools, and STAR in 15 schools. Screeners were typically administered three times a year. Students were added to and dismissed from intervention services throughout the year based on whether they had met grade-level expectations. Along with screener scores, several factors (e.g., teacher and parental input, scores on alternative tests, and need for remediation in reading) were used in the selection of students for the intervention. In the first cohort of MAF schools (schools designated as MAF schools in the first grant cycle that started in the 2015–2016 school year), the share of students in grades K–3 chosen for MAF intervention ranged from 2% to 34% (with a median of 13%) in 2015–2016 school year.[2]

In addition to universal screeners, easyCBM was used to monitor progress among intervention students. Kindergarten students took the Numbers and Operations Measurement, while students in grades 1 to 3 took the Numbers and Operations Measurement and the Numbers and Operations in Algebra Measurement. easyCBM was administered three times a year (October, January, and May), and results were reported in a centralized database.

Requirements for Teachers

The MIT must be a highly trained, certified teacher with at least 3 years of teaching experience, with preference given to teachers with at least 3 years of primary teaching experience or training in mathematics intervention services for primary students. At least one-half of the MIT's time is required to be spent delivering mathematics interventions to lower-performing primary-grade students. Administrative records show that these interventions were typically provided in small groups of four to six students with each session lasting about 30 to 45 min. Typically, five to six sessions were offered every day for 5 days a week. Importantly, these sessions did not replace regular mathematics instruction, a key requirement of the MAF grant. A review of MIT weekly schedules demonstrates that MITs scheduled interventions around homeroom teachers' schedules so that students were not pulled during core instruction. In addition to direct instruction for targeted students, the MIT spent about 5.6 hr per week coteaching mathematics with regular classroom teachers and about 3.8 hr on lesson planning.

MIT is also expected to engage in a significant number of professional development activities. Based on grantee performance reports, MITs typically participated in 8 to 12 events per year including webinars, multi-day training sessions, conferences, and courses offered by programs such as Add+VantageMR. The MIT is also expected to share instructional activities, books, and strategies obtained from these professional development opportunities with building staff during weekly team planning and professional learning communities (PLCs). MIT daily schedules suggest that the MITs spent about 2.3 hr per week on PLC activities.

In addition to MIT, MAF requires grantee schools to select two classroom teachers every year as Plus 2 teachers. Plus 2 teachers receive 10 days of professional development and attend, at the expense of the grantee schools, at least one state mathematics conference every year. Plus 2 teachers are required to be available for collaboration and coteaching with the MIT throughout the school year. Along with the MIT, Plus 2 teachers are expected to lead professional learning with additional teachers to build capacity in the building.

Mathematics Intervention Packages

Grantee schools are required to use one of the four MAF-approved primary mathematics intervention packages: Add+VantageMR® (AVMR), Assessing Math Concepts®, Do the Math®, and Math Recovery Intervention Specialist® (MRIS). These packages are used to provide interventions for targeted students, professional development for classroom teachers, and professional learning for intervention teachers. In the 2018 to 2019 school year, out of the 97 MAF grantee schools, the overwhelming majority chose to use AVMR

(76 schools) and MRIS (17 schools) to facilitate the training of intervention teachers and the delivery of intervention services.

Grantee Selection

Although we do not have details on the selection process such as the total number of applicants, application scoring rubrics, or any information on applicants who were not selected for funding, materials submitted by successful applicants shed light on the dimensions that the Kentucky Department of Education (KDE) valued most when selecting grantees. Specifically, each applicant addressed seven dimensions: identification of need, a description of the RtI framework, identification of which MAF-approved intervention package would be used, how students would be identified for targeted intervention, approaches to professional learning and sustainability, assessment and evaluation plan, and a budget.

Channels of Change

While improving the mathematics achievement of low-performing students is a key objective of MAF, program administrators and grantee performance reports indicate that the defining feature of MAF is its whole-school approach. MAF's emphasis on coteaching, professional development for classroom teachers, and the expectation that the MIT and Plus 2 teachers share what they have learned from professional development activities with other teachers in the building suggest that the program could lead to positive spillover effects on other students in the school. To capture the full impact of MAF, this study examines the schoolwide effects of MAF instead of focusing narrowly on the outcomes of students who were referred to interventions.[3]

MAF could theoretically improve student achievement in mathematics through three channels. The first channel is through increased interaction between children and their teachers. Because MAF requires that interventions not replace regular instruction, it results in a net gain in instruction time for intervention students. Empirical evidence shows that education interventions that add instruction time tend to improve student achievement significantly. For example, Taylor (2014) and Cortes et al. (2015) both used RD frameworks to study the effects of assigning students to an additional mathematics course and found positive effects on test scores in the short run. Similar evidence exists for reading, where Figlio et al. (2018) found significant benefits of additional instruction time using extended school days on student test scores.

The second channel is the delivery of more personalized support through MAF interventions in small-group settings. Small groups allow the interventionist to better adapt instruction to individual student needs, and students are thought to learn better and faster if the content is within their zone of proximal development (Guthrie, 2015; Stipek, 2002; Vygotsky, 1978). Small groups are also thought to help establish emotional and peer support quickly, which is particularly important in the context of early childhood development and learning (Mashburn et al., 2008). As we review later, early childhood mathematics interventions delivered in small-group settings have consistently produced positive improvements in student mathematics achievement (Pellegrini et al., 2018; Slavin et al., 2011; Wanzek et al., 2016).

A third feature of MAF that may help improve student mathematics achievement is setting clear expectations for what MIT should and should not do while encouraging close collaboration among teachers. This design feature protects the interventionist from distractions such as being asked to substitute for regular mathematics teachers in school. At the same time, the learning community approach to professional development has been demonstrated to improve teacher performance (Buysse et al., 2010). Close collaboration between MIT and classroom teachers also allows the interventionist to provide "just in time" support to children who need help with mathematics, and research has shown that working in a collaborative environment greatly improves teaching quality (Bransford et al., 1999).

Literature Review

There are hundreds of elementary-grade mathematics interventions and more than 10,000 studies on their efficacy. Less than 1% of those studies are considered rigorous enough to be included in several best-evidence syntheses

(Jacobse & Harskamp, 2011; Pellegrini et al., 2018; Savelsbergh et al., 2016; Slavin & Lake, 2008). However, the rigor of early mathematics intervention studies has improved in the last decade, and some patterns have started to emerge. For example, two types of interventions are found to produce positive results consistently.

The first is tutoring programs that provide one-to-one or small-group supplemental instruction, with average effect sizes of about 0.30 standard deviations in student test scores.[4] Tutoring programs typically require extensive professional development for the interventionists and the intervention tends to be intensive, meeting at least three times per week for at least 12 weeks, with each session lasting at least 30 min. The second type of promising intervention focuses on improving instructional processes (such as cooperative learning, classroom management, and motivation), with average effect sizes of around 0.25 standard deviations in student test scores.[5] All these programs pay close attention to student needs and behavior, and target instruction to student proficiency levels.[6] Similar to tutoring programs, teacher professional development is an integral part of instructional process interventions. Peer coaching and teacher collaboration in instructional planning are often featured in teacher professional development. Taken together, reviews of empirical evidence suggest that early mathematics intervention strategies that substantially affect children's daily experiences by emphasizing personalization, engagement, and motivation are likely to produce beneficial results (Pellegrini et al., 2018).

Based on this literature, the MAF design—intensive small-group supplemental instruction by full-time MITs, personalized support, teacher peer coaching, and collaboration between interventionists and regular classroom teachers—is consistent with the characteristics of successful early mathematics interventions. As such, our study complements the existing literature in this context in several important ways.

First, studies that find positive effects on student mathematics learning tend to be small, typically with no more than a few hundred students carried out in a single district. In small-scale studies, developers and experimenters are closely involved, to an extent that is often not realistic if interventions were scaled up. Our study adds to the existing literature by studying whether promising intervention features will scale successfully in a statewide program. Second, large-scale evaluations of elementary mathematics interventions have produced reasonably solid evidence on what interventions do *not* work, but it remains unclear which intervention strategies may scale successfully. This is because large-scale studies, typicallyrandomized controlled trials (RCTs) involving thousands of students across multiple districts or states have been focusing on curricula, benchmark assessments, and professional development (PD) for mathematics content and pedagogy.[7] To the best of our knowledge, there has been no evaluation of any statewide K–3 mathematics intervention programs. This is primarily because very few states have systematic, coherent early grade mathematics policies in place. Third, existing mathematics intervention research generally focuses exclusively on the mathematics achievement of students who were targeted for intervention, ignoring potential spillover effects on other students and outcomes or unintended adverse effects on student behaviors. Finally, because there are competing demands for attention to deficiencies in multiple domains such as reading and socioemotional competence, our study estimates the cost-effectiveness of MAF to inform the best way to deploy educational resources.

Data

In our impact analysis, we use student-level administrative school records obtained from KDE. These data cover all students in grades K–5 between 2011–2012 and 2018–2019 school years and include reading and mathematics scores of all tested students (in grades 3 through 8) on the Kentucky Performance Rating for Educational Progress (K-Prep) test along with demographic information on students, such as race, gender, free- or reduced-price lunch eligibility, English language learner status, exceptional/special education status, student age, and schools attended. We also observe information about student disciplinary problems (number of disciplinary referrals, number of suspensions, and the total length of suspensions), and attendance (including days absent, days present,

number of unexcused/excused tardy days, and the number of days possible).

Kentucky's administrative data also include detailed information on the interventions (including MAF) that students receive. The reporting of this information is at the student-intervention level, and it is mandatory. MAF intervention activities are populated by the mathematics team in each MAF school and checked three times a year by KDE to ensure high quality and timeliness of data entry. The intervention data include individual student-level information on the start and end date of the intervention, intervention program (e.g., AVMR and MRIS), tier of instruction, duration of the intervention (in minutes), frequency of intervention (times per week), intervention staff (the qualification level of the staff that is most directly providing the interventions services to the student), areas of student need (deficiency areas such as number sense, mathematics reasoning, and measurement). We use this information to better understand the type, duration, and intensity of the intervention students receive.

There were two MAF grant cycles between the redesign in 2015–2016 and the most recent school year we observed in our data (2018–2019 school year). The top panel in Figure 1 presents the number of "ever-MAF" schools (i.e., elementary schools that were ever designated as MAF during this time frame) by MAF status. There were 107 elementary schools in Kentucky that were designated as MAF in the first grant cycle that started in 2015–2016 school year. Of these schools, 51 remained as MAF in the second grant cycle that started in 2017–2018, whereas 56 left the MAF designation. Thirty-nine schools were newly designated as MAF in the second grant cycle.

Table 1 examines the student characteristics in pre-MAF years (between 2011–2012 and 2014–2015 school years) for first-cohort MAF schools (schools that were first designated as MAF in 2015–2016), second-cohort MAF schools (first-time designation in 2017–2018), and schools that were never designated as MAF between 2015–2016 and 2018–2019. The results suggest that the first- and second-cohort MAF schools were comparable along observed student characteristics: students in these

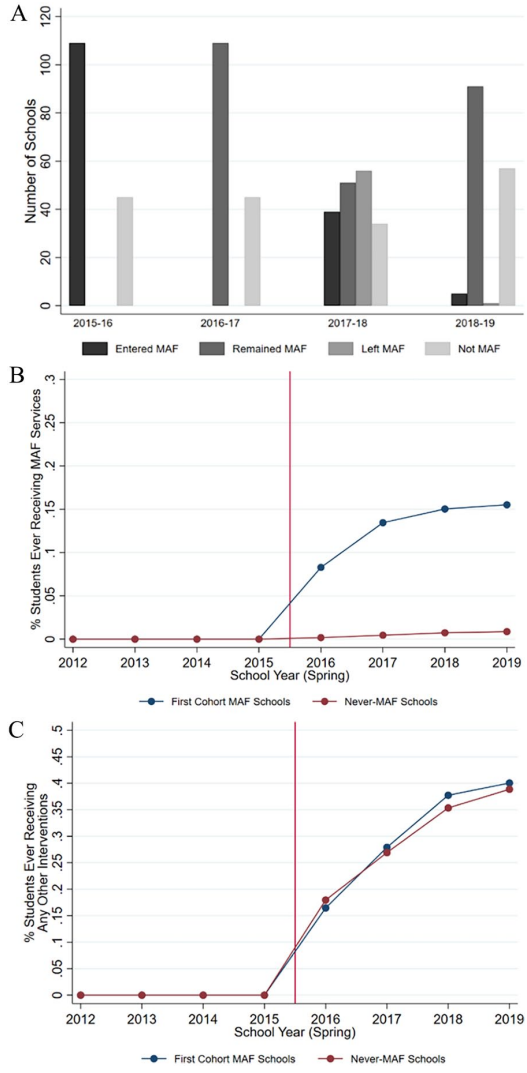


FIGURE 1. *Exposure to MAF at the school and student-level: (A) number of schools by MAF status: ever-MAF schools, (B) % of ever-MAF students: first-cohort MAF versus never-MAF schools, and (C) % of students ever identified for other services: first-cohort MAF versus never-MAF schools.*

Note. The top panel presents the number of elementary schools that were ever identified as MAF after 2015–2016 by MAF status (entered, remained, left, and not MAF) by school year between 2015–2016 and 2018–2019 school years whereas the bottom two panels present the percentage of students in first-cohort MAF schools (elementary schools that were first designated as MAF in 2015–2016 school year) and never-MAF schools who have ever received MAF services (panel B) or other services (panel C) up to (and including) that school year. MAF = Math Achievement Fund.

TABLE 1

Student Characteristics by School MAF Status, Elementary Schools Between 2011–2012 and 2014–2015 School Years

Student characteristics	School in first MAF cohort	School in second MAF cohort	Never-MAF school
Math scores (grades 3–5)	−0.019 (0.974)	−0.019 (0.960)	0.008 (1.005)
KG ready	0.351 (0.477)	0.339 (0.473)	0.356 (0.479)
% Absent days	5.052 (4.741)	4.682** (4.383)	4.641** (4.566)
Disciplinary incident	0.034 (0.181)	0.037 (0.189)	0.029 (0.167)
Received subsidized meals	0.661 (0.473)	0.655 (0.475)	0.602** (0.490)
Special education	0.113 (0.316)	0.109 (0.311)	0.103** (0.304)
English learner	0.028 (0.165)	0.026 (0.159)	0.035 (0.183)
White	0.882 (0.323)	0.887 (0.316)	0.828** (0.378)
Black	0.070 (0.255)	0.067 (0.250)	0.116*** (0.320)
Hispanic	0.058 (0.234)	0.059 (0.235)	0.064 (0.245)
Asian	0.008 (0.090)	0.008 (0.091)	0.015*** (0.121)
Foreign born	0.001 (0.034)	0.001 (0.038)	0.003*** (0.055)
Age	8.771 (1.890)	8.808 (2.026)	8.882** (1.892)
Number of unique students	80,947	32,591	436,972
Number of unique schools	107	39	720

Note. Standard deviations are given in parentheses. Math scores are standardized at the grade-year level to zero mean and unit variance. ** and *** represent that the means for the corresponding group are statistically different than the first column (students in the first MAF cohort of schools) at 5% and 1%, respectively.

schools had similar scores on math tests and were comparable along disciplinary incidents, attendance, race/ethnicity, socioeconomic status (as proxied by subsidized meal eligibility), English learner status, and special education status. That said, students in both first- and second-cohorts of MAF schools had lower math scores, were more likely to be eligible for subsidized meals, more likely to be involved in disciplinary incidents, had more absences, and were more (less) likely to be White (Black) compared to students in schools that were never designated as MAF between 2015–2016 and 2018–2019 school years.

In our cost analysis, we take a school perspective and therefore limit our data collection only to school staff who dedicated effort to the program. However, there is no formal program requirement or evidence of effort devoted on the part of volunteers or parents, leading us to believe that virtually all resources dedicated to the program are provided by schools. In turn, we use both administrative data and primary data collected through surveys administered to MITs and Plus 2 Teachers who were part of an MAF school intervention team in the 2018–2019 school year to obtain information on the school resource effort devoted to the MAF program. The survey questionnaire was organized around the activities and personnel/nonpersonnel resources associated with the MAF program.

Survey administration took place in May and June of 2021 and was sent out to the designated MIT and principal, as well as 187 Plus 2 teachers (P2Ts) from 97 MAF grantee schools that were active in the 2018–2019 school year.[8] Unfortunately, the survey data collection was conducted in the middle of the COVID-19 pandemic, which likely affected the survey response counts. Specifically, we received no responses from principals and responses from a total of 25 MITs and 8 P2Ts representing 30 schools. Among the 25 MITs who responded, 6 were dropped because they were from schools that were not included in the impact analysis.[9] Ten of the remaining 19 MITs are from schools that are also first-cohort MAF grant recipients, and 9 are from schools that received MAF grants in the second cohort only. Because the response rate among P2Ts is low, we decided to base the calculation of costs associated with these staff solely on extant administrative

information pertaining to their required participation in professional development as part of the MAF program. Also, there was only a minimal required effort on the part of principals described in the program documentation.[10] We acknowledge that by not fully capturing efforts associated with P2Ts and principals, the resulting cost estimates may be considered a lower-bound of the overall program cost. However, the costs associated with these unaccounted efforts are likely to be small.

Empirical Framework

Determining Program Effectiveness

To estimate the effects of MAF on student outcomes, we primarily use a DiD design. We estimate MAF effects for each cohort separately for two reasons. First, it is of policy interest to examine the extent to which program effects may have changed between cohorts. Program effects could change, for example, as the result of the program maturing over time; they could also change as the program scales up and the initial enthusiasm wanes. Second, we observe 4 years of post-award data for the first cohort of MAF grantee schools. This allows us to examine the extent to which program effects, if they exist, persist over time, particularly in schools that have stopped receiving MAF support.

In our main analysis, we focus on MAF schools in the first cohort, comparing the differences in student outcomes in years after versus before the designation (2015–2016 school year) with the same difference in never-MAF schools. Formally, using ordinary least squares (OLS) we estimate the following two-way fixed-effects model:[11]

$$Y_{igst} = \alpha + \beta \mathrm{MAF}_s * \delta_t + \delta_t + \mu_s + \gamma_g + \varepsilon_{igst}, \qquad (1)$$

where Y_{igst} is the outcome of interest (test scores in math and reading standardized to zero mean and unit variance at the grade-by-year level, whether the student received a suspension, and % absent days) for student i in grade g and school s in year t, MAF_s is an indicator for MAF schools in the first cohort, δ_t is year fixed-effects (with the 2014–2015 school year serving as the omitted category), and μ_s and γ_g are school and grade fixed-effects respectively. Given that MAF is a schoolwide intervention, which means there is no variation within each school/time period in treatment status among students, we cluster our standard errors at the school level. In this setting, precisely estimated zero coefficients on the interaction between MAF_s and δ_t for years prior to 2014–2015 would provide evidence that the parallel trends assumption (i.e., treatment and comparison schools would follow similar trajectories in post-treatment years in the absence of the treatment) cannot be rejected, which is a critical identification assumption in the DiD design.

Panel B in Figure 1 checks the fidelity of implementation and examines student exposure to MAF in the first-cohort MAF schools and never-MAF schools in the years before and after the MAF designation. In particular, this graph compares the share of students who ever received the MAF intervention over time between these two types of schools. The results indicate that by the end of the fourth year after the initial MAF designation, roughly 15% of all students in the first-cohort MAF schools ever received the MAF intervention compared to less than 1% in never-MAF schools (nearly all of these ever-MAF students in never-MAF schools had transferred from MAF schools).

Another concern in this context is that the first cohort of MAF schools started implementing other interventions at the same time as MAF. For example, during the time frame, we examine in this study, Kentucky also started implementing a K–3 reading intervention called Read to Achieve (RTA). If first-cohort MAF schools were more likely to implement RTA simultaneously compared to never-MAF schools, it would become harder to attribute the observed differences in student outcomes between the two types of schools in post-designation years to the causal effect of MAF. Panel C in Figure 1 repeats the same exercise in panel B, replacing MAF exposure with exposure to other interventions (including RTA). The share of students exposed to other interventions increased in both first-cohort MAF and never-MAF schools after 2014–2015 school year, but students in MAF schools were no more likely to receive another intervention compared to students in never-MAF schools.

Finally, recent developments in two-way fixed-effects models suggest that traditional DiD estimators produce biased results when the timing of treatment implementation is staggered and

TABLE 2

Key MAF Program Activities and Their Definitions Accounted for in the Cost Analysis

Program activity	Definition
Intervention services	Delivery of MAF program instruction to students, including preparing lessons, providing direct instruction, and collaborating on instruction with other teachers.
Family engagement	Engagement with parents through conferences and events.
Assessment and monitoring	Administration of assessments, identification of MAF students, and monitoring of student progress.
Professional development	Participation in professional development opportunities related to the MAF program, including formal training provided by KCM and provision of professional development and coaching by mathematics intervention teachers to other staff within schools.[a]
Administration	Planning and budgeting, scheduling service delivery, and entering data on service provision and student progress.

Note. KCM = Kentucky Center for Mathematics; MAF = Math Achievement Fund.
[a]See description of training on KCM website (https://www.kentuckymathematics.org/maf.php).

the treatment effects are heterogeneous across groups or time (Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfœuille, 2020). Separating the two cohorts of MAF grantee schools circumvents the complications related to staggered implementation. However, Sun and Abraham (2021) show that the estimated effect for year *t* could still be contaminated by both the weights and effects from periods other than year *t*. As robustness checks, we combine the two cohorts of grantees and implement the estimators developed by Callaway and Sant'Anna (2021) using both never-treated and not-yet-treated schools as the comparison. These details are presented after the main findings.

Determining Program Costs and Cost-Effectiveness

To gain a well-rounded understanding of whether the MAF program is a wise investment we cannot depend on estimates of impact alone. We must also consider the costs of implementing the program and assess these costs in relation to its effects through a cost-effectiveness analysis, which measures program cost per unit of outcome produced.

Our analysis employs the ingredients approach (Levin et al., 2018) to identify the costs of implementing the MAF program. The approach isolates the costs associated with the program by identifying the quantities of the personnel and nonpersonnel resources (ingredients) used in the implementation of the program and assigning corresponding prices to calculate their costs. Based on reviews of documentation and extant data, as well as discussions with MAF administrators, key ingredients involved in implementing MAF consist of activities related to intervention services, family engagement, assessment and monitoring, professional development, and administration (Table 2). School staff (MITs and P2Ts) surveys were used to quantify each ingredient. For personnel components, we asked staff to provide hours spent for a given program activity along with auxiliary costs related to activities in question, such as transportation and lodging/food costs for attending required formal professional development. We also collected data on staff years of experience and the highest level of education attainment to apply appropriate compensation rates in the next step. Data for nonpersonnel resources such as software and equipment utilized during MAF-related activities, we asked respondents about the use of commonly used equipment and software.

The next step in the ingredients approach is to assign prices to the resources. The prices for personnel resources are compensation rates (salary plus benefits), which vary by years of experience and the highest level of education attainment. Importantly, the personnel compensation rates used reflect statewide averages so that any variation in the subsequent costs across sites reflect differences in the qualifications of staff used and not the influence of

local labor markets on the price of staff (for more details see Supplemental Appendix A in the online version of the journal). The unit prices of nonpersonnel resources involved in the MAF program were derived from information posted by major online retailers (e.g., BestBuy, Amazon, and HP).

The final step to calculate costs simply involves multiplying the quantity of each resource used by its corresponding unit price and summing across the resources. The final calculations of the overall program implementation cost for each school sum together the costs of personnel and annualized costs of nonpersonnel resources. As resource allocation data was only collected for the 2018–2019 school year, the costs for this year are translated into present values for each of the previous 3 years and summed to provide a 4-year implementation cost for each school (in 2018–2019 dollars) and expressed in per-pupil terms (more details can be found in Supplemental Appendix A in the online version of the journal). We calculate an enrollment-weighted average of the 4-year program per-pupil costs across the cohort 1 grantee schools (using school K–5 enrollment as the weight), which serves as the key cost metric for the cost-effectiveness analysis.

This 4-year weighted average per-student cost is then coupled with the estimated 4-year MAF impacts on student outcomes to generate cost-effectiveness ratios. Each ratio shows the cost per unit of outcome produced by the intervention. For example, the cost-effectiveness of the MAF program with respect to outcome (o) is simply the average per-student cost ($Average\ Cost\ Per\ Student_{MAF}$) divided by the impact (μ_o), as follows:

$$\text{Cost-Effectiveness}_o = \frac{\text{Average Cost Per Student}_{\text{MAF}}}{\mu_o} \quad (2)$$

Results

Effectiveness of MAF

Figure 2 presents the event study estimates (along with their 95% confidence intervals) for our four main outcomes of interest: math (panel A) and reading (panel B) scores on K-Prep in grades 3 through 5; whether the student was involved in a disciplinary incident in panel C and percent absent days in panel D (both estimated separately for grades K–5 and K–3). Table 3 presents these coefficients, estimated without (in columns labeled as [I]) and with (columns labeled as [II]) student covariates (i.e., an indicator for subsidized meal receipt, race/ethnicity, gender, special education status, English learner status, foreign-born indicator, and age).

The main takeaway from this analysis is that MAF had a significant positive effect on both test and non-test outcomes of students, especially beyond the first year after the designation. In particular, while we do not find any concerning evidence of differential pre-treatment trends between first-cohort MAF and never-MAF schools in the outcomes of interest, we find significant differences in the years after MAF designation. For example, MAF increased student mathematics test scores by 0.05 standard deviation after 2 years, 0.08 standard deviation after 3 years, and 0.09 standard deviation after 4 years. Similar, yet slightly smaller effects emerge for reading scores, with second-year effects of 0.03 standard deviation, third-year effects of 0.04 standard deviation, and fourth-year effects of 0.06 standard deviation. The gradually increasing effects of MAF are in line with (a) more students who received the MAF intervention in grades K–3 being tested in grades 3 through 5 and (b) the gradual implementation of effective practices regarding math instruction throughout the school.

We also find that MAF significantly reduced disciplinary incidents and student absences in participating schools. For example, MAF decreased the likelihood that students were involved in disciplinary incidents by 1 percentage point in the first year (by 25% of the dependent variable mean), 2 percentage points in the second year (by roughly 50%), and 3 percentage points in the third year (by 75%).[12] Similarly, MAF led to a decline in the percent absent days of roughly 0.1 to 0.3 percentage points in the first 4 years (about 2%–6% of the dependent variable mean).[13] Panels C and D in Figure 2 also present the results for students in earlier grades (K–3) to see whether the effects were larger in the grades directly targeted by MAF interventions. The results suggest that the K–3 effects are mostly in

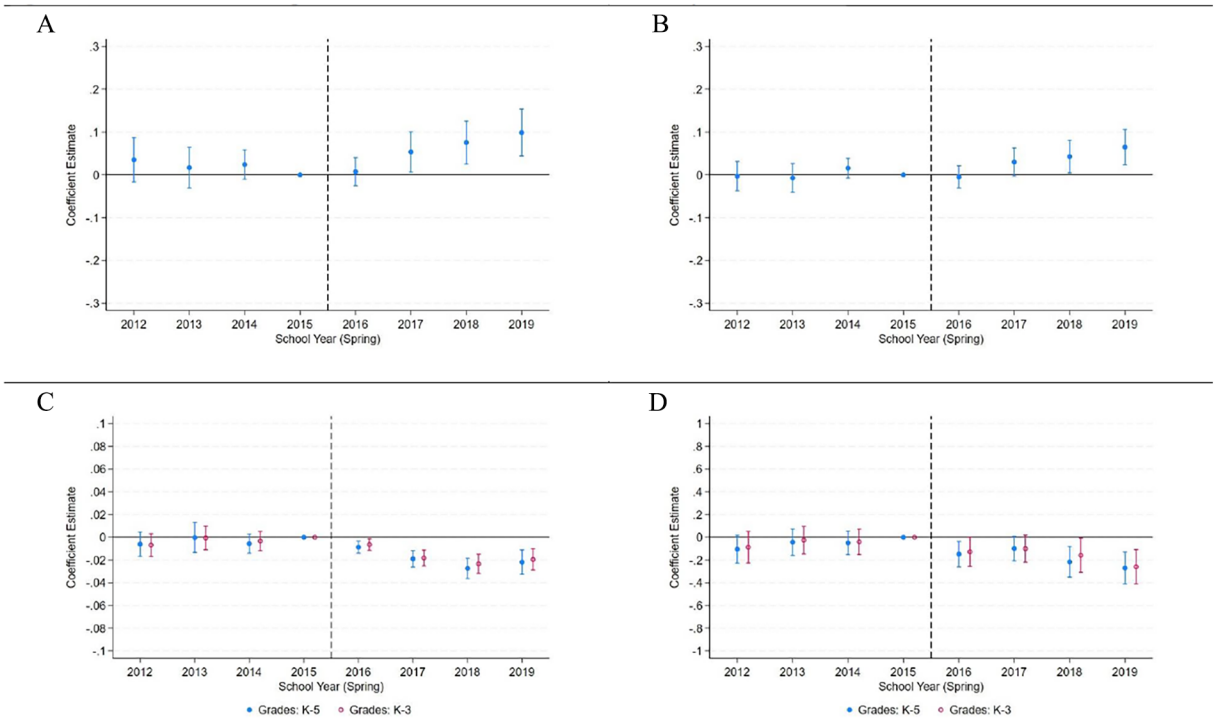


FIGURE 2. *Effects of MAF designation on student outcomes: event study estimates. (A) math scores, (B) reading score, (C) disciplinary incidents, and (D) % absent days.*

Note: Each dot presents the estimated coefficient on the interaction term between the MAF indicator and the indicator for the corresponding school year, with the spikes providing the 95% confidence intervals. The estimates are obtained using schools in the first cohort of MAF schools as the treatment and never-MAF schools as the comparison group with school, grade, and year fixed effects. Standard errors are clustered at the school level.

MAF = Math Achievement Fund.

TABLE 3

Effects of MAF Designation on Student Outcomes: Event Study Estimates

Treatment: first-cohort MAF schools, comparison: never-MAF

Year indicators and covariates	Math scores		Reading scores		Disciplinary incidents		% Absent days	
	(I)	(II)	(I)	(II)	(I)	(II)	(I)	(II)
2011–2012 SY	0.035	0.030	−0.003	−0.008	−0.006	−0.006	−0.105*	−0.102*
	(0.026)	(0.025)	(0.017)	(0.017)	(0.005)	(0.005)	(0.062)	(0.062)
2012–2013 SY	0.017	0.013	−0.007	−0.011	−0.000	−0.000	−0.043	−0.036
	(0.024)	(0.023)	(0.017)	(0.016)	(0.007)	(0.007)	(0.059)	(0.058)
2013–2014 SY	0.024	0.023	0.016	0.015	−0.006	−0.005	−0.049	−0.043
	(0.017)	(0.017)	(0.012)	(0.012)	(0.004)	(0.004)	(0.053)	(0.051)
2015–2016 SY	0.007	0.004	−0.005	−0.008	−0.009***	−0.009***	−0.148***	−0.137**
	(0.017)	(0.017)	(0.013)	(0.013)	(0.003)	(0.003)	(0.057)	(0.058)
2016–2017 SY	0.053**	0.053**	0.030*	0.030*	−0.019***	−0.019***	−0.099*	−0.093*
	(0.024)	(0.024)	(0.017)	(0.017)	(0.004)	(0.004)	(0.055)	(0.055)
2017–2018 SY	0.075***	0.075***	0.043**	0.041**	−0.027***	−0.027***	−0.216***	−0.215***
	(0.026)	(0.025)	(0.019)	(0.019)	(0.005)	(0.005)	(0.069)	(0.068)
2018–2019 SY	0.099***	0.093***	0.065***	0.057***	−0.022***	−0.022***	−0.271***	−0.273***
	(0.028)	(0.028)	(0.021)	(0.021)	(0.005)	(0.005)	(0.072)	(0.072)
Mean of Y	0.005	0.005	0.006	0.007	0.039	0.040	4.755	4.753
N	1,123,838	1,120,885	1,123,838	1,120,884	2,383,336	2,368,845	2,351,178	2,337,473
School Fixed Effects (FE)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student covariates	No	Yes	No	Yes	No	Yes	No	Yes

Note: Robust standard errors, clustered at the school level, are given in parentheses. The estimates represent the estimated coefficient on the interaction term given with 2014–2015 school year (the year before the policy took effect) serving as the baseline category. Student covariates include an indicator for subsidized meal receipt, race/ethnicity, gender, special education status, English learner status, foreign-born indicator, and age. *, **, and *** represent statistical significance at 10%, 5%, and 1%, respectively. MAF = Math Achievement Fund.

line with the overall effects on non-test outcomes, which provides suggestive evidence that the effects of MAF designation go above and beyond the students chosen for the intervention.

Our findings also indicate that the MAF effects persist even after schools leave the program, which provides further evidence that systemic, schoolwide shifts in MAF schools are an important driver behind the observed benefits. In particular, Figure 3 breaks down the analysis in Figure 2 by whether the school remained as MAF in the second grant cycle (that started in 2017–2018 school year) or left the MAF program. The estimated MAF effects are comparable for the two types of schools in the first 2 years of their designation (2015–2016 and 2016–2017 school years). What is more interesting is that we find significant benefits of the program on student outcomes in the third and fourth years after their designation when they no longer receive MAF support from the state. For example, for these schools, we find MAF effects of 0.1 standard deviations in math and significant reductions in disciplinary incidents and absences in the fourth year.

Do certain student groups benefit more from MAF? Gender, for example, is widely documented to be associated with mathematics performance and STEM (Science, Technology, Engineering, and Math) choices (Card & Payne, 2021; Ellison & Swanson, 2010). Gender differences emerge in the early years of elementary school (Penner & Paret, 2008), and MAF may have differential impacts on boys and girls. In addition, research suggests that low-income and

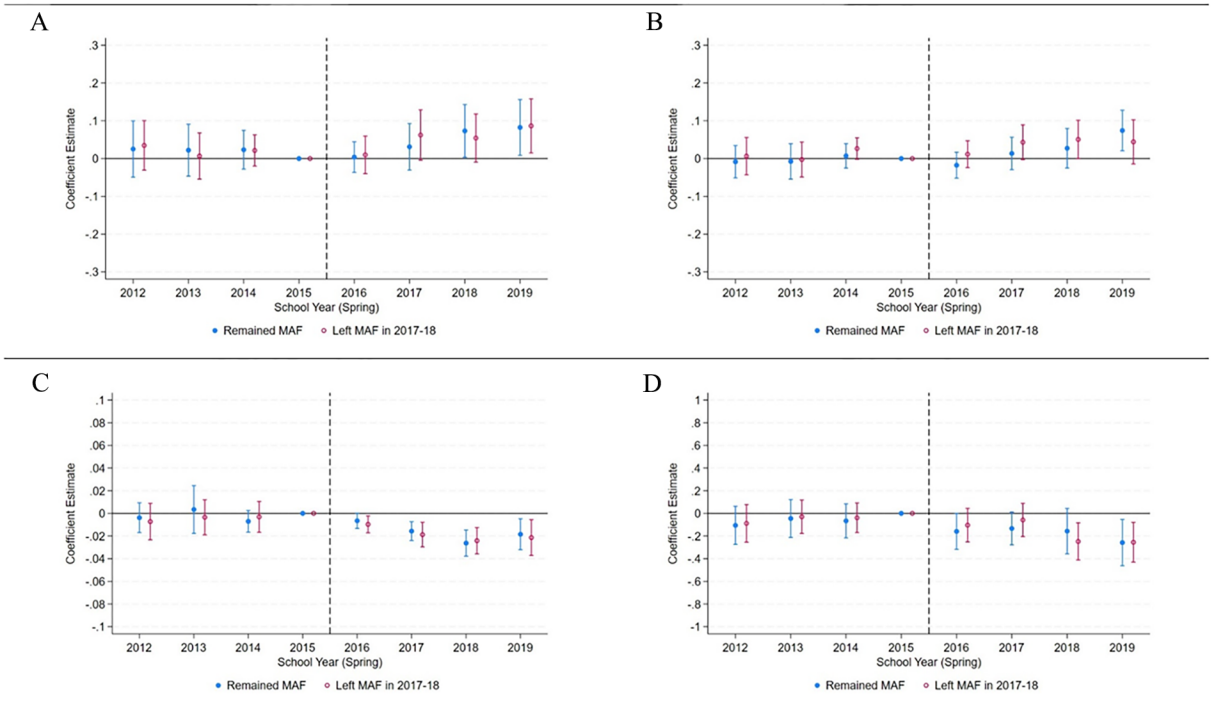


FIGURE 3. *Effects of MAF designation on student outcomes: event study estimates, by school duration in MAF: (A) math scores, (B) reading score, (C) disciplinary incidents, and (D) % absent days.*

Note: Each dot presents the estimated coefficient on the interaction term between the MAF indicator and the indicator for the corresponding school year, with the spikes providing the 95% confidence intervals. The estimates are obtained using schools in the first cohort of MAF schools that left MAF at the end of 2017–2018 and those that remained as MAF until 2018–2019 as the treatment groups and never-MAF schools as the comparison group with school, grade, and year fixed-effects. Standard errors are clustered at the school level. MAF = Math Achievement Fund.

racially minority students tend to be more susceptible to being labeled as in need of interventions (Mokher et al., 2018; Ou, 2010; Papay et al., 2011). Finally, one possible unintended consequence of receiving MAF grants is that resources may be reallocated within school to focus more on grades K–3.

Figure 4 to 7 examines this question and breaks down the analysis in Figure 2 by student grade (in tested grades) in Figure 4, by whether the student received subsidized meals in Figure 5, by student gender in Figure 6, and by student race/ethnicity in Figure 7. Table 4 presents the traditional, static DiD estimates (after MAF versus before MAF, in first-cohort MAF schools vs. never-MAF schools) obtained using the outcomes of interest in Figure 2 (with K–5 incidents and absence rates) overall and by student gender, subsidized meal receipt, and race/ethnicity. This table also presents the DiD results using an "ever-MAF" indicator that equals 1 if the student ever received the MAF intervention as the outcome to assess whether the MAF designation has a differential effect on the likelihood of receiving the intervention for different student groups.

The overarching conclusion from this analysis is that the benefits of MAF are widespread. In particular, we do not find any differences in MAF effects on outcomes of interest across grades, by student gender, or by student subsidized meal receipt although we find that MAF designation has a significantly larger effect on ever receiving the intervention for students who receive subsidized meals (an effect of 15 percentage points vs. 8 for students who do not receive subsidized meals).

That said, the lack of overlap in confidence intervals suggests significant differences between White and non-White (mostly Black and Hispanic students). While MAF designation does not have a differential effect on the likelihood of receiving the intervention for these two student groups, the benefits on test scores and non-test outcomes are significantly larger for racial minorities. For example, we find MAF effects of 0.13 standard deviation in math and 0.10 standard deviation in reading for racial minorities compared to no statistically significant effects for White students. MAF led to declines in disciplinary incidents and absence rates among both racial minority and White students, but the decline in disciplinary incidents is significantly larger among racial minorities (6 percentage points, equivalent to 67% of the dependent variable mean for the student group) than among White students (0.7 percentage points, equivalent to 26% of the dependent variable mean). The MAF effects on absence rates are not statistically different between racial minorities and White students (a decline of 0.28 percentage points for racial minorities compared to 0.10 percentage points for White students).

What about differential effects by the share of students receiving the MAF intervention? For example, if the observed MAF effect is primarily driven by the effect of the intervention on treated students, then one would expect the effects to be larger in schools with a larger share of students receiving the intervention. The primary challenge in this analysis is that the number of students who directly receive intervention services under MAF in a school is endogenous and it depends partly on the effectiveness of the intervention in prior years (i.e., an effective intervention in the prior year would improve students' mathematics skills in the current year and hence reducing the share of students receiving intervention services). Therefore, in Figure 8 we break down the analysis in Figure 2 by the share of K–3 students who received direct interventions in first-cohort MAF schools in 2015–2016 school year (first year of implementation). We find no significant differences in MAF effects for MAF schools with higher (above median) and lower (below median) share of K–3 students identified for MAF in the first year of implementation. This finding once again provides evidence that the policy impacts student outcomes in ways beyond its effect on students directly receiving intervention services.

An alternative explanation for the observed MAF effect is the possible effect of MAF designation on student composition. For example, if the designation leads to lower-performing students leaving these schools or higher-performing students entering, it could lead to improved outcomes in MAF schools on average compared to other schools. In Supplemental Appendix Figure 1 (available in the online version of this article) examines this possibility and checks the effects of MAF designation on observed student characteristics (kindergarten readiness for students in

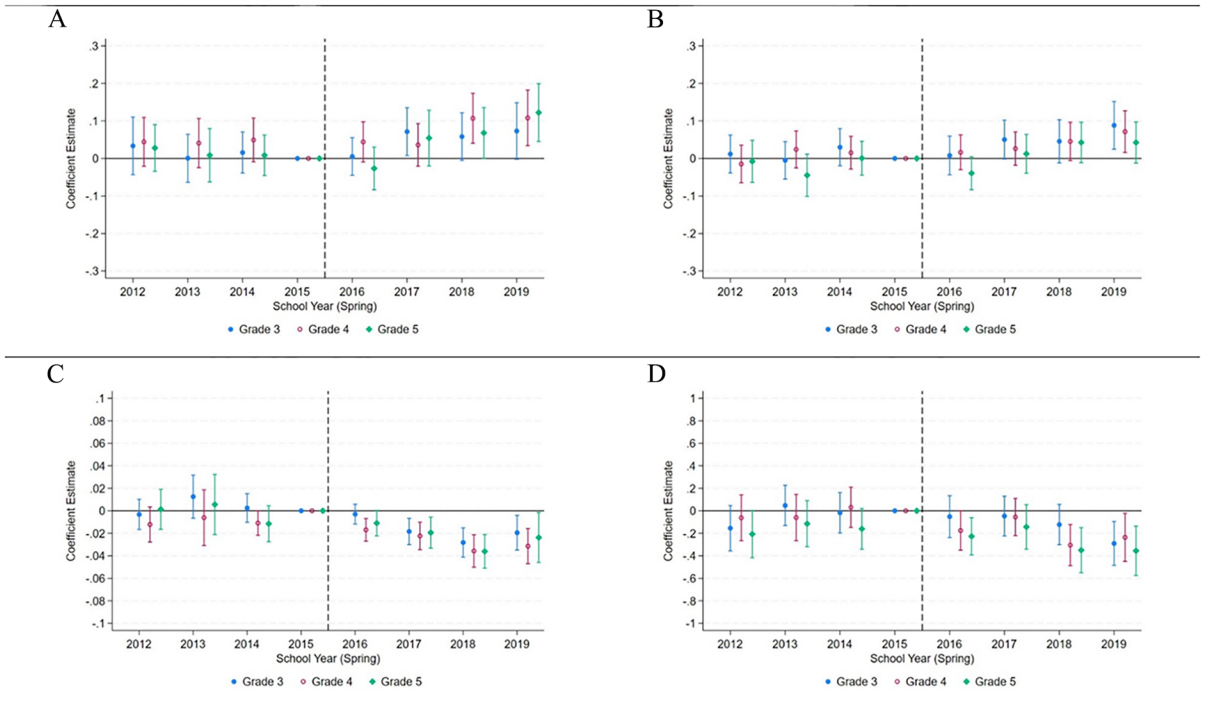


FIGURE 4. *Effects of MAF designation on student outcomes: event study estimates, by grade: (A) math scores, (B) reading score, (C) disciplinary incidents, and (D) % absent days.*

Note: Each dot presents the estimated coefficient on the interaction term between the MAF indicator and the indicator for the corresponding school year, with the spikes providing the 95% confidence intervals. The estimates are obtained using schools in the first cohort of MAF schools as the treatment and never-MAF schools as the comparison group with school, grade, and year fixed-effects. Standard errors are clustered at the school level. MAF = Math Achievement Fund.

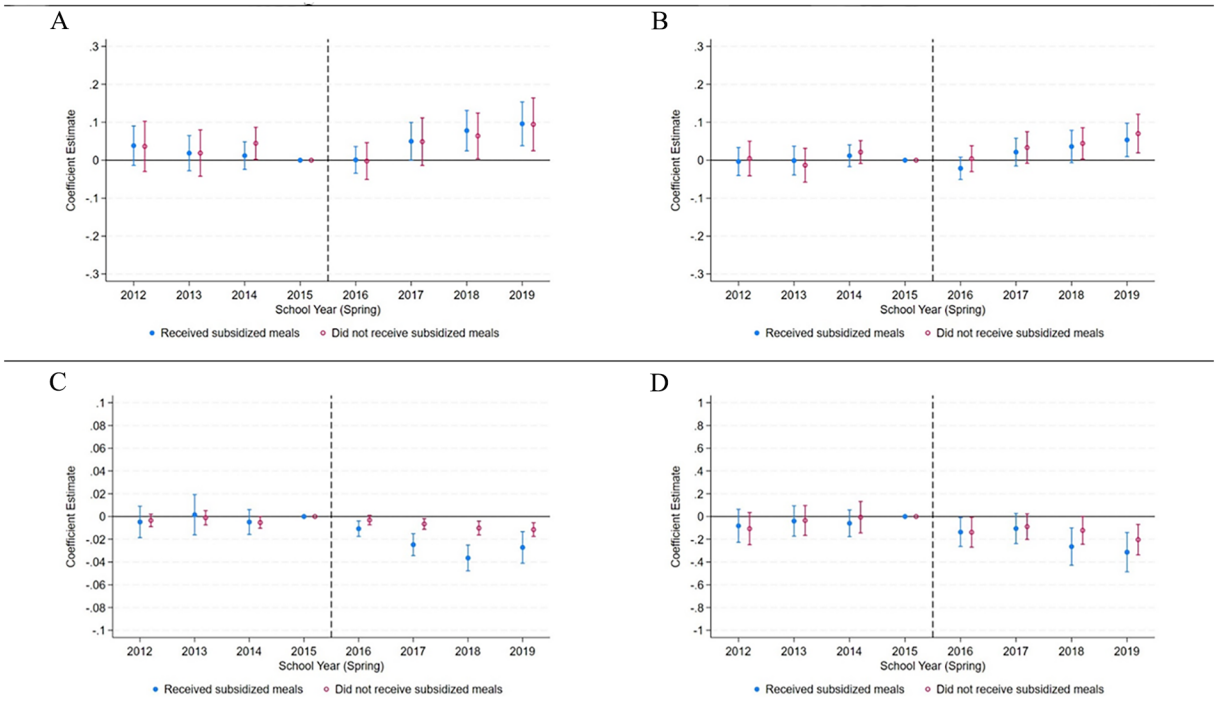


FIGURE 5. *Effects of MAF designation on student outcomes: event study estimates, by student subsidized meal receipt: (A) math scores, (B) reading score, (C) disciplinary incidents, and (D) % absent days.*

Note: Each dot presents the estimated coefficient on the interaction term between the MAF indicator and the indicator for the corresponding school year, with the spikes providing the 95% confidence intervals. The estimates are obtained using schools in the first cohort of MAF schools as the treatment and never-MAF schools as the comparison group with school, grade, and year fixed-effects. Standard errors are clustered at the school level. MAF = Math Achievement Fund.

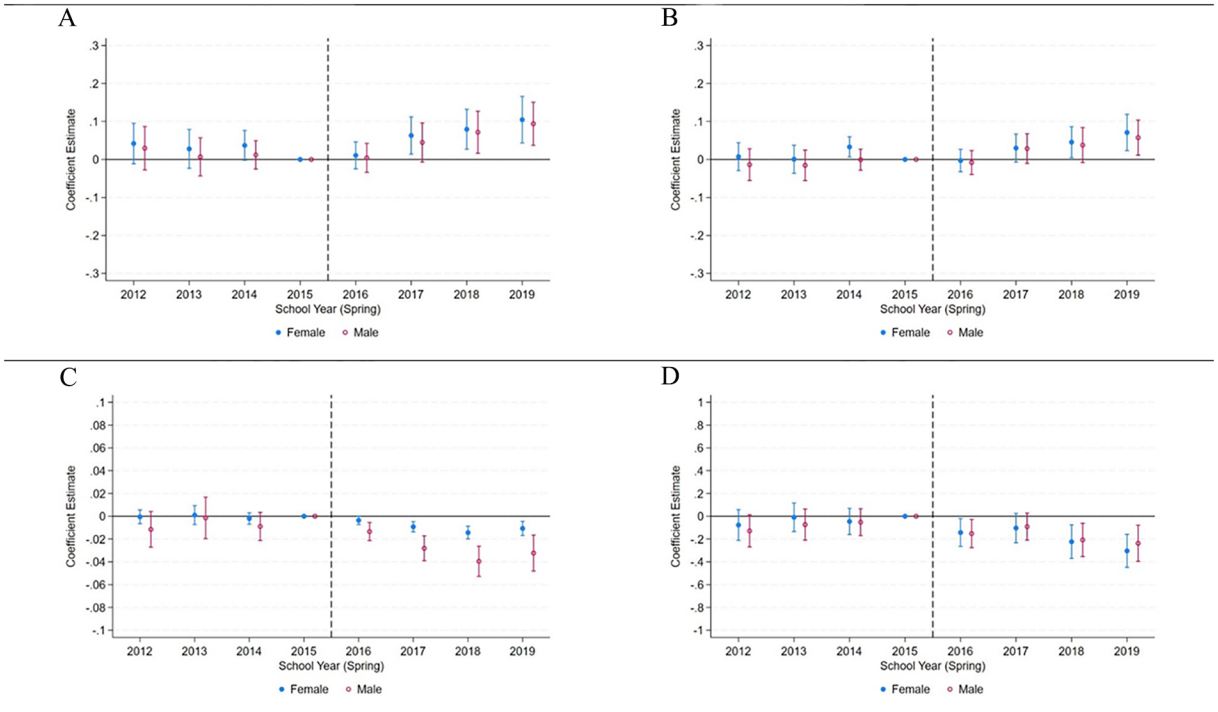


FIGURE 6. *Effects of MAF designation on student outcomes: event study estimates, by student gender: (A) math scores, (B) reading score, (C) disciplinary incidents, and (D) % absent days.*

Note: Each dot presents the estimated coefficient on the interaction term between the MAF indicator and the indicator for the corresponding school year, with the spikes providing the 95% confidence intervals. The estimates are obtained using schools in the first cohort of MAF schools as the treatment and never-MAF schools as the comparison group with school, grade, and year fixed-effects. Standard errors are clustered at the school level. MAF = Math Achievement Fund.

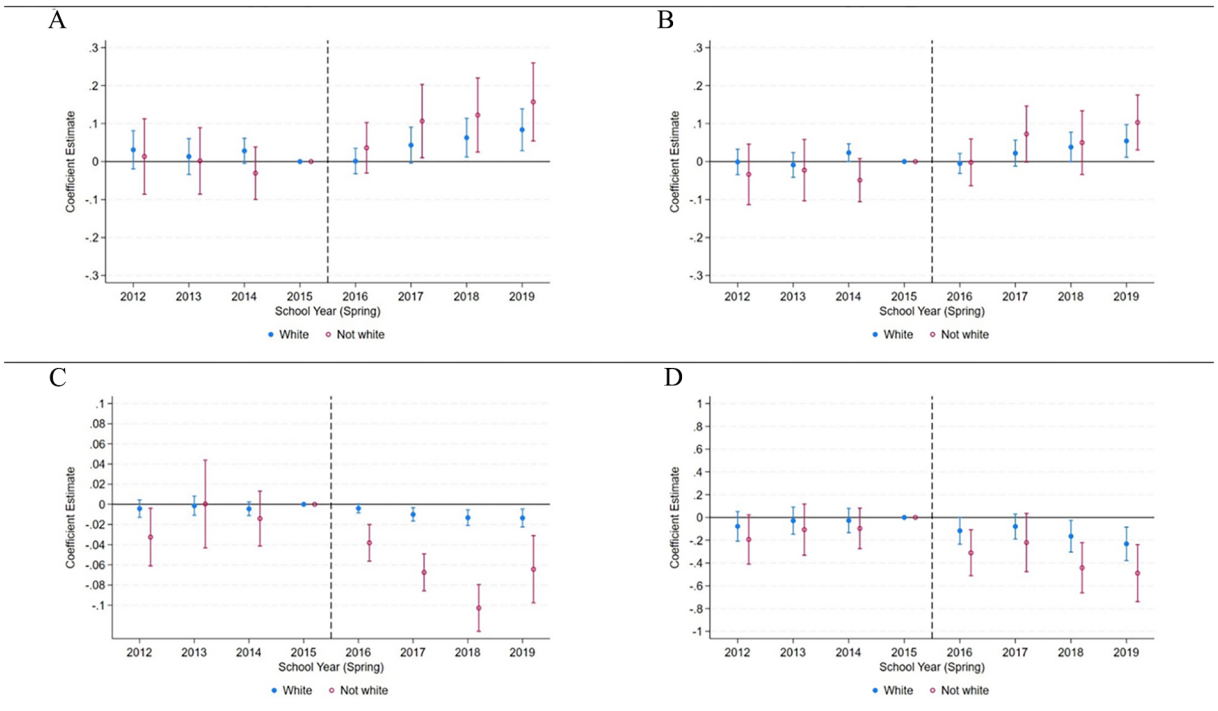


FIGURE 7. *Effects of MAF designation on student outcomes: event study estimates, by student race/ethnicity: (A) math scores, (B) reading score, (C) disciplinary incidents, and (D) % absent days.*

Note: Each dot presents the estimated coefficient on the interaction term between the MAF indicator and the indicator for the corresponding school year, with the spikes providing the 95% confidence intervals. The estimates are obtained using schools in the first cohort of MAF schools as the treatment and never-MAF schools as the comparison group with school, grade, and year fixed-effects. Standard errors are clustered at the school level. MAF = Math Achievement Fund.

TABLE 4

Effects of MAF Designation on Receiving MAF Services and Student Outcomes, DiD Estimates, Overall and by Subgroup

Treatment: first cohort MAF schools, Comparison: never-MAF

Outcomes	All	Female	Male	Subsidized meals	No subsidized meals	White	Non-White
Ever received MAF services	0.125***	0.134***	0.117***	0.147***	0.078***	0.125***	0.128***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.008)
90% CI	[0.118, 0.133]	[0.125, 0.143]	[0.109, 0.125]	[0.138, 0.155]	[0.071, 0.086]	[0.117, 0.133]	[0.115, 0.142]
Mean of Y	0.012	0.013	0.011	0.016	0.006	0.013	0.009
Math scores	0.040**	0.036*	0.044**	0.043**	0.025	0.025	0.128***
	(0.020)	(0.020)	(0.021)	(0.020)	(0.026)	(0.020)	(0.034)
90% CI	[0.007, 0.072]	[0.003, 0.068]	[0.009, 0.078]	[0.010, 0.077]	[−0.017, 0.068]	[−0.008, 0.059]	[0.079, 0.185]
Mean of Y	0.005	0.006	0.005	−0.242	0.414	0.068	−0.312
Reading scores	0.031*	0.022	0.037**	0.024	0.032*	0.019	0.104***
	(0.016)	(0.017)	(0.017)	(0.018)	(0.019)	(0.016)	(0.035)
90% CI	[0.005, 0.057]	[−0.005, 0.050]	[0.009, −0.066]	[−0.005, 0.053]	[0.001, 0.064]	[−0.007, 0.046]	[0.046, 0.162]
Mean of Y	0.007	0.082	−0.065	−0.231	0.401	0.081	−0.363
Disciplinary incident	−0.016***	−0.009***	−0.023***	−0.023***	−0.005***	−0.007***	−0.058***
	(0.003)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.009)
90% CI	[−0.021, −0.011]	[−0.012, −0.006]	[−0.029, −0.016]	[−0.029, −0.017]	[−0.008, −0.002]	[−0.011, −0.004]	[−0.074, −0.043]
Mean of Y	0.040	0.017	0.060	0.054	0.015	0.031	0.087
% Absent days	−0.133***	−0.161***	−0.105*s	−0.158***	−0.104**	−0.098**	−0.275***
	(0.049)	(0.050)	(0.054)	(0.060)	(0.045)	(0.050)	(0.085)
90% CI	[−0.214, −0.052]	[−0.244, −0.078]	[−0.194, −0.017]	[−0.257, −0.060]	[−0.179, −0.029]	[−0.181, −0.016]	[−0.415, −0.135]
Mean of Y	4.753	4.759	4.747	5.494	3.474	4.822	4.392
N (tested grades)	1,120,885	549,092	571,789	698,038	422,847	935,134	185,751
N (all grades)	2,368,845	1,149,404	1,219,418	1,486,883	881,962	1,986,014	382,831

Note: Robust standard errors, clustered at the school level, are given in parentheses. The estimates represent the estimated coefficient on the interaction term given with 2014–2015 school year (the year before the policy took effect) serving as the baseline category. All regressions control for school, year, and grade fixed- effects, and the student covariates, including an indicator for subsidized meal receipt, race/ethnicity, gender, special education status, English learner status, foreign-born indicator, and age. *, **, and *** represent statistical significance at 10%, 5%, and 1%, respectively. MAF = Math Achievement Fund.

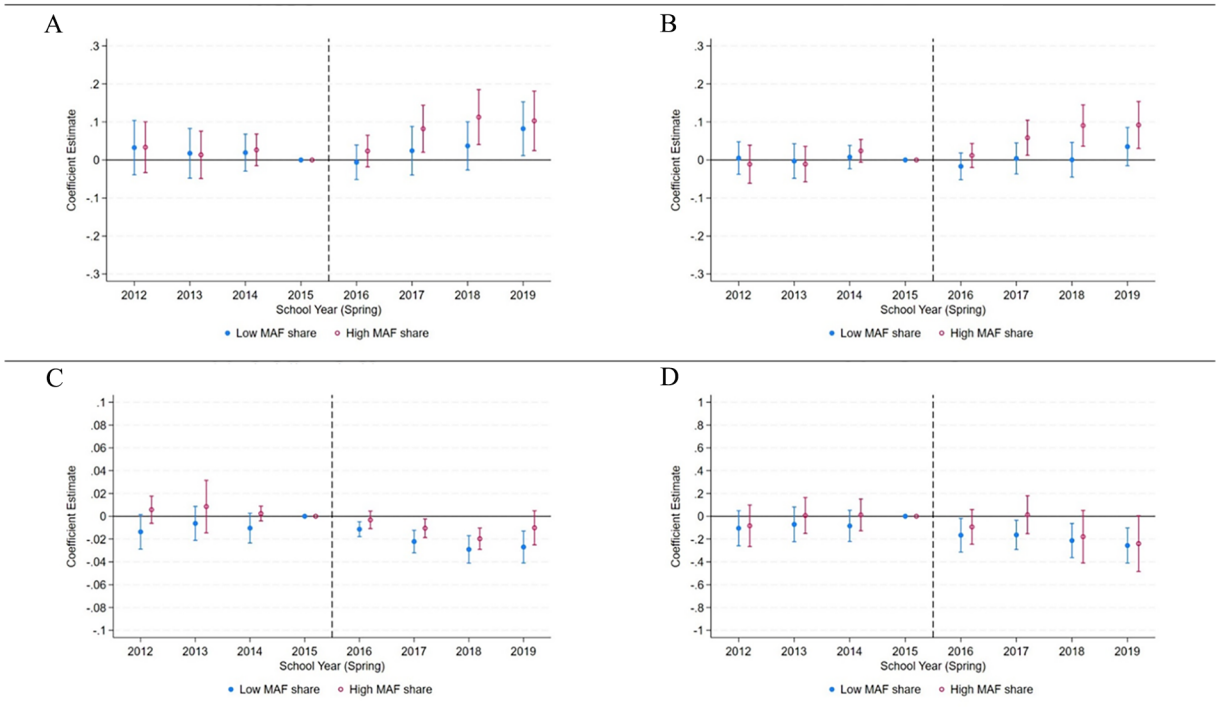


FIGURE 8. *Effects of MAF designation on student outcomes: event study estimates, by the share of K–3 students receiving MAF intervention in 2015–2016 in first-cohort MAF schools: (A) math scores, (B) reading score, (C) disciplinary incidents, and (D) % absent days.*

Note: Each dot presents the estimated coefficient on the interaction term between the MAF indicator and the indicator for the corresponding school year, with the spikes providing the 95% confidence intervals. The estimates are obtained using schools in the first cohort of MAF schools as the treatment and never-MAF schools as the comparison group with school, grade, and year fixed-effects. Standard errors are clustered at the school level. MAF = Math Achievement Fund.

kindergarten, whether the student receives subsidized meals, race/ethnicity, special education status, English learner status, gender, and immigrant status). Overall, we do not find evidence of shifts in student composition large enough to explain the observed effects of MAF on student outcomes. This is consistent with the finding presented in Table 3 that shows that the estimated effects of MAF remain virtually unchanged when we include student covariates in the regressions.

Effects for the Second MAF Cohort

How do these estimated compare with the effects on the second cohort of MAF schools? Supplemental Appendix Figure 2 (available in the online version of this article) repeats the analysis in panels B and C in Figure 1 for the schools that were first designated as MAF in 2017 to 2018 school year and Supplemental Appendix Figure 3 (available in the online version of this article) presents the event study estimates obtained using these schools as the treatment group and never-MAF schools as the comparison group. Similar to the first cohort, MAF designation increases exposure to the MAF intervention, and we find no significant differences in exposure to other interventions between first-time second-cohort MAF schools and never-MAF schools.

In terms of MAF effects on student outcomes, we find results that are somewhat in line with the findings using the first cohort of MAF schools in the first 2 years after their designation. We find no significant effects on math scores and absence rates after 2 years although it is important to note that these coefficients are less precisely estimated as we have fewer treatment schools in this exercise and there is a slight upward trend in MAF effects on math scores in years after the designation. In contrast, we find significant benefits of MAF designation on reading scores and disciplinary incidents in the second year. In all cases, we find no concerning differences in pretreatment trends between treatment and comparison groups.

Robustness Checks

In this section, we examine (a) the overall MAF effects across both grantee cohorts and (b) the robustness of the main findings to alternative comparison schools. Combining both MAF cohorts entails staggered treatment timing. Research demonstrates that traditional DiD estimates are biased (and often with little direct policy relevance) when treatment timing is staggered and treatment effects vary across groups and time periods (de Chaisemartin & D'Haultfœuille, 2020; Goodman-Bacon, 2021). Here, we follow the empirical strategy developed by Callaway and Sant'Anna (2021) that is robust to heterogeneous treatment effects in a staggered setting and estimate an event study regression of the following form:

$$Y_{igst} = \delta_t + \mu_s + \gamma_g + \sum_{l=-K, l\neq -1}^{L} \beta_l 1\{l_s = t - g_s\} + \varepsilon_{igst}. \quad (3)$$

Similar to Equation 1, Y_{igst} is the outcome of interest for student i in grade g and school s in year t. δ_t, μ_S and γ_g are year, school, and grade fixed-effects. Because implementation timing now varies by grantee cohort, the term $\text{MAF}_s * \delta_t$ in Equation 1 is replaced by a series of relative time binary variables $1\{l_s = t - g_s\}$ that represent the l th year since school s first received MAF grants in year g_s. Thus, $l = 0$ for the year in which a school first received MAF. As is conventional and consistent with Equation 1, the year immediately before MAF designation ($l = -1$) is the reference period and is omitted. In other words, the treatment-control difference in outcomes for each period $l \neq -1$ is compared with the treatment-control difference in ($l = -1$). When $l \geq 0$, β_l estimates the cumulative effect $l+1$ years after initial MAF designation. When $l < -1$, β_l estimates the placebo effect $|l|$ periods before MAF designation.

This event study model was estimated using two alternative sets of comparison schools: Never-MAF schools and not-yet-MAF schools. Our main analysis has so far used never-MAF schools as the comparison. However, those schools could be inherently different from schools that applied and received MAF grants in ways that are unobservable to researchers. As a result, not-yet-MAF schools were used as an alternative comparison group with the plausible assumption that these schools likely resembled schools that had received MAF grants earlier more closely than never-MAF schools. With only two cohorts of MAF schools, using not-yet-MAF

schools as the comparison in essence uses cohort 2 MAF schools as the control group for cohort 1 MAF schools.

These estimates are reported in Table 5 as Column III (never-MAF as the comparison) and IV (not-yet-MAF as the comparison). The primary DiD estimates for cohort 1 and cohort 2 are reproduced in Columns I and II for comparison purposes. Results obtained using Callaway and Sant'Anna estimators are highly consistent with the main findings. For math, the Callaway-Sant'Anna estimates are 0.01, 0.05, 0.06, and 0.08 standard deviation for the first four post-MAF years, respectively, when never-MAF schools are the comparison. These are nearly identical to the main DiD results reported for Cohort 1 MAF schools (that is, 0.00, 0.05, 0.08, and 0.09 for year t, $t + 1$, $t + 2$, and $t + 3$, respectively). The coefficients are statistically significant at the end of the second, third, and fourth years since MAF was first implemented. When not-yet-MAF schools are the comparison, the estimate effects are limited to the first two post-MAF years, and they are 0.02 and 0.06 standard deviations. These are also in line with the main results, albeit statistically insignificant as the sample size is only 15% as large as the sample size used for the main results. Similarly, the estimated effects for other outcomes are robust to choices of samples, comparison groups, and estimators.

Cost-Effectiveness Estimates

The following section reports the results of both the cost and cost-effectiveness analyses. The findings provided for the main cost analysis are based on those first-cohort MAF schools that were two-time grant recipients and therefore implemented the program for 4 years from 2015–2016 to 2018–2019. The findings for the cost-effectiveness analysis are based on all schools in the main impact analysis sample, which includes both 2- and 4-year grant recipients who received their initial MAF grants in 2015–2016. As reported earlier, MAF impact estimates are very similar between grantees schools that received grants for 2 and 4 years (see Figure 3).

The overall 4-year per-pupil cost of MAF is $750 (Figure 9). As is the case with most education programs, most of the cost ($726 or 96%) is attributable to personnel resources. The breakout of costs by program activities in Figure 10 shows that over half of the overall program cost ($431 or 57%) is dedicated to providing direct intervention services to low-performing students, while just over a quarter ($194 or 26%) is spent on professional development for program teachers. Smaller shares of the program cost are associated with assessing students and monitoring their progress ($72 or 10%), program administration ($25 or 3%), and engaging families ($28 or 4%).

Figure 11 shows considerable variation in program cost across the grantee schools. We divide MAF grantee schools into terciles based on their per-pupil costs. It depicts the average per-pupil costs for low-, medium-, and high-cost MAF schools, as well as how these costs break out across the program activities.[14] The average program cost per pupil for MAF schools in the top tercile is $1,062, nearly twice as much as the per-pupil cost for MAF schools in the bottom tercile ($604). The cost breakout shows that the additional cost associated with the high-cost group can be attributed to greater spending across all activities, especially on intervention services and administration.

Figure 12 illustrates how per-pupil program costs in 2018–2019 vary with the scale at which MAF schools operate.[15] It shows a clear pattern consistent with economies of scale where the per-student cost of schools operating with a smaller enrollment tends to be higher than those with larger enrollments. For instance, the schools with fewer than 300 K–5 students have annual per-pupil costs of around $350, while the per-pupil costs of those with more than 600 students are close to $150.

By combining the cost estimates and the estimated math achievement impact we find that the MAF program has a cost-effectiveness ratio of $8,069 in terms of the targeted outcome of interest, student achievement in mathematics. The cost-effectiveness ratio is an estimate, and therefore it is bounded by a confidence interval within which the true value lies with some degree of certainty. Using the Monte Carlo interval method documented in Dong et al. (2024), we find that the 95%-confidence interval of the estimated cost-effectiveness ratio to be between $5,038 and $19,735.[16]

TABLE 5

Effects of MAF Designation on Student Outcomes: Robustness Checks

Year relative to initial implementation (t)	Math scores				Reading scores			
	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)
$t-6$		0.046 (0.050)	−0.034* (0.020)			0.021 (0.030)	−0.019 (0.019)	
$t-5$		0.019 (0.048)	−0.003 (0.025)			0.012 (0.031)	−0.006 (0.024)	
$t-4$	0.030 (0.025)	0.010 (0.036)	−0.009 (0.012)	0.016 (0.023)	−0.008 (0.017)	−0.006 (0.027)	−0.002 (0.010)	0.011 (0.020)
$t-3$	0.013 (0.023)	0.015 (0.036)	0.005 (0.015)	0.018 (0.029)	−0.011 (0.016)	0.009 (0.023)	0.018 (0.013)	0.037 (0.027)
$t-2$	0.023 (0.017)	−0.001 (0.025)	−0.015 (0.014)	−0.041* (0.024)	0.015 (0.012)	−0.002 (0.016)	−0.011 (0.010)	−0.036 (0.022)
t	0.004 (0.017)	0.031 (0.026)	0.005 (0.014)	0.017 (0.025)	−0.008 (0.013)	0.021 (0.019)	−0.004 (0.011)	0.006 (0.021)
$t+1$	0.053** (0.024)	0.046 (0.035)	0.045** (0.020)	0.064 (0.041)	0.030* (0.017)	0.044** (0.022)	0.027** (0.013)	0.040 (0.027)
$t+2$	0.075*** (0.025)		0.064** (0.025)		0.041** (0.019)		0.033* (0.018)	
$t+3$	0.093*** (0.028)		0.082*** (0.028)		0.057*** (0.021)		0.047** (0.020)	
N	1,120,885	1,021,085	1,179,373	162,253	1,120,884	1,021,083	1,179,371	162,252
Pre-trend test (χ^2)			7.814	3.527			9.026	4.424
p-Value			0.452	0.317			0.340	0.219
Year relative to initial implementation (t)	Disciplinary incidents				% Absent days			
	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)
$t-6$		0.002 (0.007)	0.011 (0.009)			−0.038 (0.104)	0.054 (0.073)	
$t-5$		0.013 (0.008)	0.000 (0.008)			0.023 (0.104)	0.060 (0.077)	
$t-4$	−0.006 (0.005)	0.013* (0.007)	0.002 (0.004)	−0.005 (0.010)	−0.102* (0.062)	0.095 (0.096)	0.023 (0.046)	−0.027 (0.084)
$t-3$	−0.000 (0.007)	0.006 (0.006)	−0.002 (0.004)	−0.005 (0.010)	−0.036 (0.058)	−0.037 (0.084)	−0.007 (0.045)	−0.045 (0.080)
$t-2$	−0.005 (0.004)	0.011*** (0.004)	0.001 (0.003)	0.011 (0.008)	−0.043 (0.051)	−0.045 (0.080)	0.112 (0.072)	0.161* (0.087)
t	−0.009*** (0.003)	−0.009* (0.005)	−0.007*** (0.003)	−0.014*** (0.005)	−0.137** (0.058)	−0.037 (0.084)	−0.156** (0.077)	−0.135 (0.083)
$t+1$	−0.019*** (0.004)	−0.010** (0.006)	−0.014*** (0.003)	−0.014** (0.006)	−0.093* (0.055)	−0.097 (0.090)	−0.165** (0.075)	−0.144 (0.097)
$t+2$	−0.027*** (0.005)		−0.022*** (0.005)		−0.215*** (0.068)		−0.300*** (0.105)	
$t+3$	−0.022*** (0.005)		−0.018*** (0.006)		−0.273*** (0.072)		−0.421*** (0.114)	
N	2,368,845	2,153,398	2,501,438	364,741	2,337,473	2,122,664	2,469,065	362,101
Pre-trend test (χ^2)			13.218	2.548			6.462	3.551
p-Value			0.105	0.467			0.596	0.314

Note. Robust standard errors, clustered at the school level, are given in parentheses. The year immediately before the policy took effect ($t-1$) serves as the baseline category. All models include school, year, and grade fixed effects and student covariates including an indicator for subsidized meal receipt, race/ethnicity, gender, special education status, English learner status, foreign-born indicator, and age. Column (I) includes cohort 1 only. Column (II) includes cohort 2 only. Columns (III) and (IV) include both cohorts, with never-MAF schools and not-yet-MAF schools as the comparison group, respectively. *, **, and *** represent statistical significance at 10%, 5%, and 1%, respectively. MAF = Math Achievement Fund.

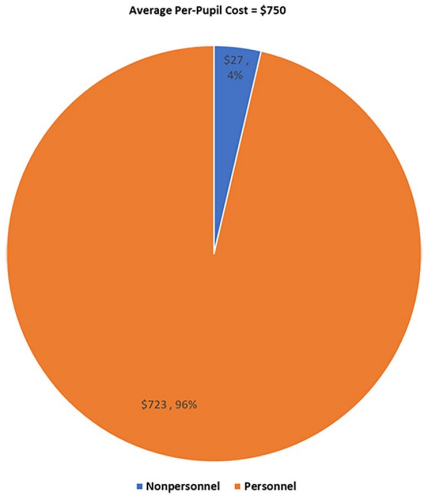


FIGURE 9. *Average MAF cost per K–5 pupil by personnel and nonpersonnel for MAF implementation schools (in 2018–2019 dollars).*
Note. MAF = Math Achievement Fund.

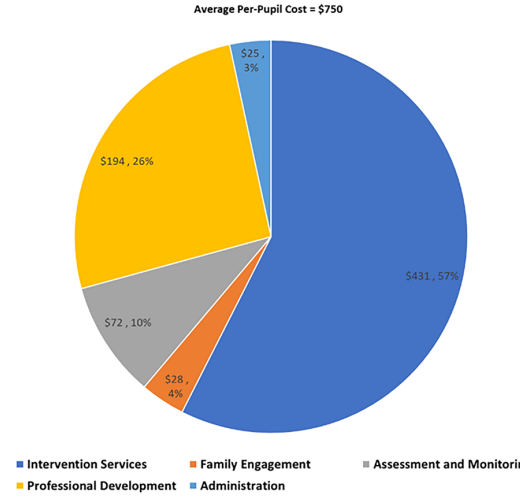


FIGURE 10. *Average MAF cost per K–5 pupil by program activity for MAF implementation schools (in 2018–2019 dollars).*
Note. MAF = Math Achievement Fund.

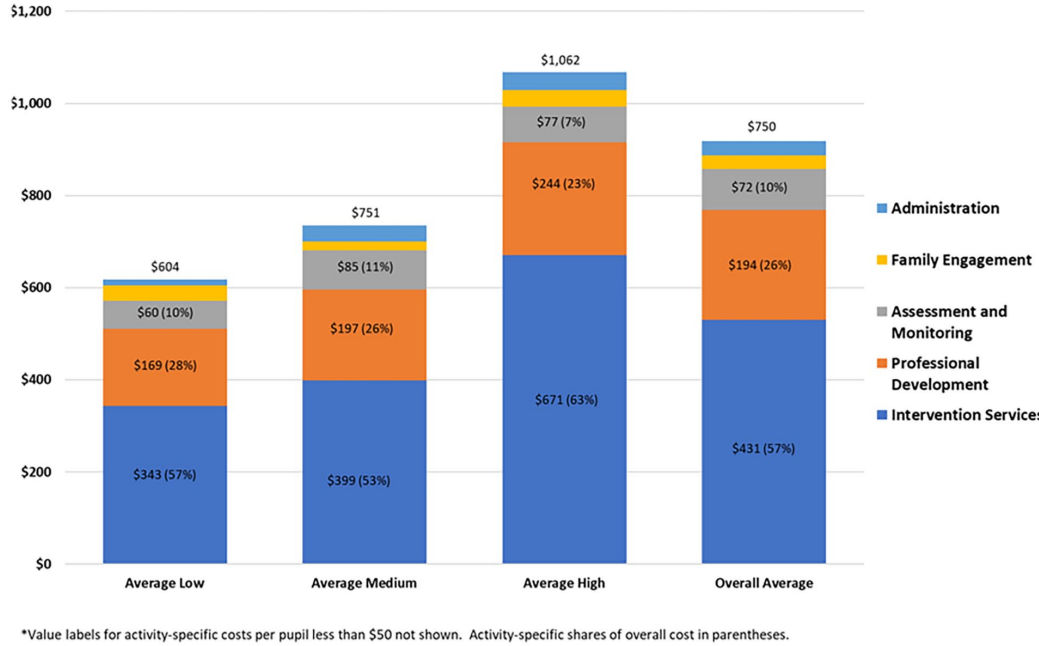


FIGURE 11. *Average MAF program cost per pupil by activity in high, medium, and low-cost MAF implementation schools (in 2018–2019 dollars).*
Note. MAF = Math Achievement Fund.

We also searched the literature for comparable cost-effectiveness findings on interventions aimed at improving student math achievement outcomes. Table 6 reports cost-effectiveness ratios on math achievement scores from alternative interventions reported in Yeh (2010), Foster et al. (2013), and Barrett and VanDerHeyden (2020), which have been inflation-adjusted to reflect 2018–2019

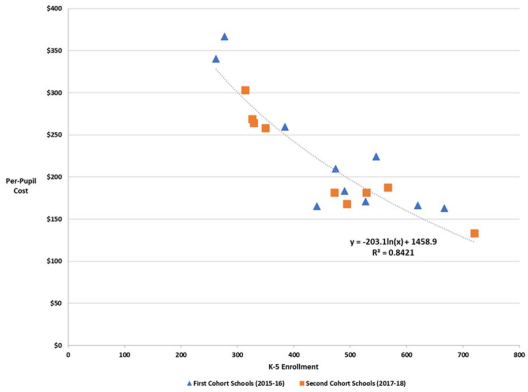


FIGURE 12. *Relationship between annual program cost per K–5 pupil and K–5 enrollment for MAF implementation schools (in 2018–2019 dollars).*
Note. MAF = Math Achievement Fund.

dollars so that they can be compared to the MAF program cost-effectiveness ratio reported for math achievement. The comparison reveals that the MAF program CER estimate is less cost-effective than programs aimed at improving teacher quality such as the Appalachian Math and Science Partnership (Foster et al., 2013) and teachers obtaining a master's degree (Yeh, 2010). While the point estimate of the MAF program suggests that it is more cost-effective than alternative programs aimed at improving classroom quality such as offering more rigorous math classes (Yeh, 2010) and class size reduction (Finn et al., 2001; Nye et al., 2001), the ratios for these programs fall within the MAF 95% confidence interval denoted by the shaded rows in Table 6 suggesting that their cost-effectiveness does not differ statistically from that of MAF. Note that without estimates of the confidence intervals for the other interventions with lower cost-effectiveness ratios, it is impossible to determine whether they differ statistically from that of the MAF program.

Concluding Remarks

Children starting their K–12 education today are more diverse in their knowledge and skills than before (Diamond et al., 2013), and K–12 schools are called upon to make up for deficiencies in foundational competencies that many entering children display due to their preschool experiences. Gaps in early grade mathematics proficiency are stubbornly difficult to close, and early delays in mathematical development lead to lower growth rates in mathematics skills later. The MIT from one of the grantee schools wrote:

> The reason we wrote the grant originally was because of our fall to fall data. It always showed no growth, and actually seemed to increase in need as the students got older. We were doing a great job of working really hard to reduce the number of tier 3 students from fall to spring, but only to have an increase in the number of students needing math interventions steadily climb as students got older.

Other MITs shared similar frustrations. How a state should design a coherent program that remedies and prevents mathematics development delays before they become intractable is an urgent question for state education policymakers to consider.

Findings from this study confirm that key elements of successful early mathematics interventions—small-group instruction, personalized support for students, peer coaching, and close collaboration among teachers—can be scaled up moderately without losing effectiveness. To put the size of the MAF effect into perspective, the fourth-year effect on mathematics achievement is equivalent to a class size reduction of four to five students (Angrist & Lavy, 1999) or replacing all novice teachers with teachers with 3 to 5 years of experience (Xu, Özek, & Hansen, 2015). But at a per-pupil cost of $750 for 4 years of intervention, MAF could be considerably cheaper than these other policy alternatives.

Importantly, the MAF benefits were found across grade levels and sustained even after grantee schools stopped receiving MAF support. One possible explanation for the sustained impact is the combination of targeted support for K–3 students and attention to improving mathematics instruction in regular classrooms. Researchers have pointed out that mathematics interventions—even when successful—must be paired with classroom practices to maintain the gains in mathematics skills achieved during the interventions (Clements et al., 2013; Smith et al., 2013; Watts et al., 2017). Because of the critical role that post-intervention learning environments play, it is essential to improve regular classroom teaching even if the sole objective of a program is to help students who are struggling with mathematics. In addition, all MAF grantees committed

TABLE 6

Relative Cost-Effectiveness of Interventions on Math Achievement

Interventions	Effect sizes (standard deviation)	Cost per student (in 2019)	CER (cost per 1 standard deviation)
Classwide math intervention[a]	0.18	$255	$1,415
AMSP training[b]	0.03	$57	$1,889
Obtaining a master's degree for teachers[c]	0.22	$898	$4,082
MAF—Monte Carlo lower bound$_{95\% CI}$	NA	NA	$5,038
MAF—Point estimate	0.09	$750	$8,069
Students taking more rigorous math courses[c]	0.20	$2,443	$12,215
Class size reduction, TN STAR Project 1[d]	0.13	$1,763	$13,667
Class size reduction, TN STAR Project 2[e]	0.09	$1,763	$19,589
MAF—Monte Carlo upper bound$_{95\% CI}$	NA	NA	$19,735

Note. Shaded rows represent upper and lower bounds of 95% confidence interval around MAF program cost-effectiveness ratio point estimate. AMSP = Appalachian Math and Science Partnership; MAF = Math Achievement Fund.
[a]Barrett and VanDerHayden (2020).
[b]Foster et al. (2013).
[c]Yeh (2010).
[d]Finn et al. (2001).
[e]Nye et al. (2001).

in their applications to continued support for at least some aspects of the program even after the funding expired. For example, many grantees proposed to use funds from Title I, extended school services, and the general fund to continue the provision of mathematics intervention packages and to support professional learning. A few grantees also planned to keep the role of MIT after the MAF grant expiration.

MAF produced similar benefits for all students regardless of sex, subsidized meal status, or race/ethnicity. For some outcomes, the impact is larger for students who are eligible for subsidized meals and for racial minority students. These findings, if replicable in more schools and states, may hold the key to addressing inequalities in later life outcomes. This is because efforts to address achievement gaps in secondary schools tend to be too little, too late.[17] Some of the proposed policy options are likely met with more challenges in practice than early grade interventions.[18]

The extent to which a program like MAF can be further scaled up remains an open question. The success of MAF, in our view, is partially attributable to the management of the program. With just over 90 grantee schools (fewer than 1/10 of all primary schools in Kentucky), MAF was closely managed by an experienced mathematics teacher at KDE. A centralized database was used to collect and monitor the number of intervention students and their progress multiple times a year. Individual MITs were contacted whenever discrepancies or data anomalies occurred. Before each school year, proposed daily schedules for all MITs were collected and approved. At the end of the school year, surveys were administered to collect feedback from MITs about their teaching and learning experiences. At the end of each grant cycle, grantees that wished to renew were evaluated using a 14-point rubric. This level of involvement will be challenging when more schools are involved, and the quality of management will likely vary when multiple program managers are needed.

Even at the current scale, grantee schools reported several implementation challenges. These include difficulty in recruiting qualified MITs, turnovers of MIT and Plus 2 teachers, insufficient number of training slots, and difficulty in creating an intervention schedule that did not conflict with any core classes. These issues are likely to become more acute if more schools were to be included in programs like MAF.

In addition to how well programs such as MAF could be further scaled up, future research should also investigate how accurately existing K–3 mathematics screeners reflect and predict mathematics skills development. Some MITs

reported concerns that students identified for mathematics intervention might have a deficiency in language instead of mathematic skills. Although we find no research evidence for early grade mathematics screeners, a study on interim assessments used for reading screening in North Carolina concluded that its K–2 screeners did not adequately identify students at risk of scoring below proficient on the state reading assessment at the end of grade 3 (Koon et al., 2020). There is a need for more research so that districts can make informed choices of early grade mathematics screeners that are based on their reliability, validity, and classification accuracy (Petscher et al., 2019).

Acknowledgments

The authors would like to thank their research partners Karen Dodd, Erin Chavez, Aaron Butler, and Hannah Poquette from the Kentucky Department of Education; Barrett Ross from the Kentucky Center for Statistics; Kelly DeLong from the Kentucky Center for Mathematics; and Mary Lee Glore from Northern Kentucky University. The authors would also like to thank the participants of the SREE conference for valuable feedback. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Kentucky Department of Education, the Kentucky Center for Statistics, the Kentucky Center for Mathematics, or Northern Kentucky University.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the National Science Foundation under Grant No. 2000483.

ORCID iDs

Zeyu Xu https://orcid.org/0000-0003-4894-0141
Umut Özek https://orcid.org/0000-0003-2790-9612
Jesse Levin https://orcid.org/0000-0003-0965-0790
Dong Hoon Lee https://orcid.org/0000-0003-1521-133X

Notes

1. Although nearly half of all states have specific requirements for early literacy assessments and interventions, only three states at the time of writing had policies that mandate additional mathematics support for K–3 students based on screening test results (Georgia Kindergarten Inventory of Developing Skills, Maine's pilot Numeracy4ME, and Kentucky's Mathematics Achievement Fund, which is the focus of the current study). A handful of other states require numeracy screening before grade 3, but these states do not stipulate interventions based on test results.

2. MAF excludes students with Individualized Educational Plans (IEPs) related to math goals. Students with IEPs receive specialized instruction from special education teachers.

3. Ideally, we would also want to estimate the extent to which MAF improves the outcomes of targeted students. The use of universal screeners in grantee schools lends support to a potential regression discontinuity design. Unfortunately, an examination of program implementation indicates that screener scores were one of the many inputs used in intervention referrals. An empirical analysis of screener scores confirms that the designated cutoffs on universal screeners are not binding, and there is no significant discontinuity in treatment probability at the cutoffs.

4. Examples of tutoring programs include *Mathematics Recovery* (Smith et al., 2013), *Galaxy Mathematics* (Fuchs et al., 2013), *Pirate Mathematics* (Fuchs et al., 2010), *Number Rockets* (Gersten et al., 2015), and *FocusMATH* (Styers & Baird-Wilkerson, 2011).

5. Examples of instructional process programs include *Team Assisted Individualization (TAI)* (Stevens & Slavin, 1995; Slavin & Karweit, 1985), *PAX Good Behavior Game* (Weis et al., 2015), and *Individualizing Student Instruction* (Connor et al., 2018).

6. By contrast, interventions that focus on mathematics curricula (textbooks), benchmark assessment and teacher professional development for mathematics knowledge or pedagogy have generally produced little effect on early grade mathematics learning (Pellegrini et al., 2018). For example, *Mathematics Solutions* and *Cognitively Guided Instructions*, two widely used mathematics professional development programs that focus on improving teachers' mathematics knowledge and pedagogy, are found to have no discernible effect on student mathematics learning (Jacob et al., 2017; Schoen et al., 2018). Mathematics curricula are found to have no detectable relationship with fourth- and fifth-grade students' mathematics achievement growth (Blazar et al., 2019).

7. These studies generally show null program effects (Garet et al., 2016; Newman et al., 2012; Randel et al., 2016).

8. Not all the second-cohort schools were eligible for the analysis. Out of the 97 schools, 4 schools joined in the second year of the grant cycle, 4 schools did not serve students in grades 3 to 5 and had no outcome data.

9. Three of the six schools only serve grades K through 2 and therefore have no state standardized test scores, two schools received off-cycle grants (i.e., they joined MAF in the second year of the second MAF grant cycle), and one school does not have complete historical student performance data that are needed to estimate the MAF impact.

10. Specifically, program documentation does not detail any effort on the part of principals outside of a single 6-hr training session on mathematics and cognitive coaching that occurs in the second year of the grant cycle.

11. Examples of studies that use repeated cross-sectional data at the individual level to estimate group-level effects include Bertrand et al. (2004) and Rouse et al. (2013).

12. It is important to note that this decline could be driven by two factors: (a) change in student behavior and/or (b) increased leniency among teachers and school administrators for student misbehavior after MAF designation.

13. There may be concerns that multiple comparisons will increase the likelihood of false positives. In Supplemental Appendix Tables 2 and 3 (available in the online version of this article), we present the *p*-values that correspond to the comparisons presented in Tables 3 and 4, respectively, along with *p*-values adjusted for comparing multiple outcomes using two methods: sharpened false discover rate *q*-values and Bonferroni correction. The *q*-values are the preferred adjustment because Bonferroni is known to increase the probability of false negatives (Anderson, 2008; Menyhart et al., 2021).

14. Specifically, we calculated the K–5 enrollment-weighted averages of per-pupil program costs for the seven least costly, six most costly schools, and six medium-cost MAF schools from which resource allocation data was collected.

15. The 19 schools are composed of 10 first-cohort MAF schools and 9 second-cohort MAF schools. The 9 second-cohort schools were plotted in Figure 12 because we included them in the imputation model for the 89 first-cohort schools with missing cost data. However, the 9 second-cohort MAF schools were not included in the impact analysis sample, which consists only of the first-cohort schools (2015–2016 school year).

16. Estimation of the 95% confidence interval bounding the point estimate of the MAF program mathematics achievement cost-effectiveness was facilitated by the tool developed as part of the study by Dong et al. (2024).

17. Transition interventions, for example, are designed to help secondary school students get ready for college. Even though 39 states have offered transition intervention programs as of 2017, empirical evidence suggests that they had no effect on helping students succeed in college.

18. For example, academic tracking (e.g., only advanced mathematics students can take algebra 2 in eighth grade) is shown to be related to within-school segregation of racial/ethnic groups (Clotfelter et al., 2021). But tracking, especially in mathematics, reflects differential skills accumulation over many years. Policies that call for opening access to advanced secondary mathematics courses to all students raise concerns about instructional challenges, diminished rigor, and extra workload for underprepared students to simultaneously learn advanced topics and close preexisting gaps in mathematics skills. Addressing gaps in mathematics skills early appears to be a more sensible policy choice.

References

Altonji, J. G., & Pierret, C. R. (2001). Employer learning and statistical discrimination. *Quarterly Journal of Economics*, *116*(1), 313–350.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575.

Austin, W., Figlio, D., Goldhaber, D., Hanushek, E. A., Kilbride, T., Koedel, C., Lee, J. S., Luo, J., Özek, U., Parsons, E., Rivkin, S. G., Sass, T. R., & Strunk, K. O. (2023). Academic mobility in US public schools: Evidence from nearly 3 million students. *Journal of Public Economics*, *228*, 105016.

Barrett, C., & VanDerHeyden, A. (2020). A cost-effectiveness analysis of classwide math intervention. *Journal of School Psychology*, *80*, 54–65.

Baumert, J., Nagy, G., & Lehmann, R. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools?. *Child Development*, *83*(4), 1347–1367.

Behrman, J. R., Ross, D. R., & Sabot, R. (2008). Improving quality versus increasing the quantity of schooling: Estimates of rates of return from rural Pakistan. *Journal of Development Economics*, *85*(1–2), 94–104.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. *The Quarterly Journal of Economics*, *119*(1), 249–275.

Blazar, D., Heller, B., Kane, T., Polikoff, M., Staiger, D., Carrell, S., Goldhaber, D., Harris, D. N., Hitch, R., Holden, K. L., & Kurlaender, M. (2019). *Learning by the book: Comparing math achievement growth by textbook in six common core states* (Research report). Center for Education Policy Research, Harvard University.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. National Academy Press.

Buysse, V., Castro, D. C., & Peisner-Feinberg, E. (2010). Effects of a professional development program on classroom practices and outcomes for Latino Dual Language Learners. *Early Childhood Research Quarterly*, *25*(1), 94–206.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Card, D., & Payne, A. A. (2021). High school choices and the gender gap in STEM. *Economic Inquiry*, *59*(1), 9–28.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812–850.

Clotfelter, C. T., Ladd, H. F., Clifton, C. R., & Turaeva, M. R. (2021). School segregation at the classroom level in a southern "New destination" state. *Race and Social Problems*, *13*(2), 131–160.

Connor, C. M., Mazzocco, M. M., Kurz, T., Crowe, E. C., Tighe, E. L., Wood, T. S., & Morrison, F. J. (2018). Using assessment to individualize early mathematics instruction. *Journal of School Psychology*, *66*, 97–113.

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment long-run impacts of double-dose algebra. *Journal of Human Resources*, *50*(1), 108–158.

Cross, C. T., Woods, T. A., & Schweingruber, H. E. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. National Academies Press.

de Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2696.

Diamond, K. E., Justice, L. M., Siegler, R. S., & Snyder, P. A. (2013). *Synthesis of IES research on early intervention and early childhood education* (NCSER 2013-3001). National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Dong, N., Maynard, R., Kelcey, B., Spybrook, J., Li, W., Bowden, A. B., & Pham, D. (2024). Advantages of Monte Carlo confidence intervals for incremental cost-effectiveness ratios: A comparison of five methods. *Journal of Research on Educational Effectiveness*, 1–29. https://doi.org/10.1080/19345747.2024.2393412

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., Japel, C., & Sexton, H. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428.

Ellison, G., & Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American Mathematics Competitions. *Journal of Economic Perspectives*, *24*(2), 109–128.

Figlio, D., Holden, K. L., & Ozek, U. (2018). Do students benefit from longer school days? Regression discontinuity evidence from Florida's additional hour of literacy instruction. *Economics of Education Review*, *67*, 171–183.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., DeSelms, J., Seethaler, P. M., Wilson, J., Craddock, C. F., Bryant, J. D., Luther, K., & Changas, P. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology*, *105*(1), 58–77.

Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., & Hamlett, C. L. (2010). The effects of strategic counting instruction, with and without deliberate practice, on number combination skill among students with mathematics difficulties. *Learning and Individual Differences*, *20*(2), 89–100.

Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (2001). The enduring effects of small classes. *Teachers College Record*, *103*, 145–183.

Foster, J., Toma, E., & Troske, S. (2013). Does teacher professional development improve Math and Science outcomes and is it cost effective? *Journal of Education Finance*, *38*(3), 255–275.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., & Borman, G. D. (2016).

Focusing on mathematical knowledge: The impact of content-intensive teacher professional development (NCEE 2016-4010). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, H. D. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLOS ONE*, *8*(1), e54651.

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge large-scale replication of a randomized controlled trial. *American Educational Research Journal*, *52*(3), 516–546.

Glewwe, P. (1996). The relevance of standard estimates of rates of return to schooling for education policy: A critical assessment. *Journal of Development Economics*, *51*(2), 267–290.

Goldhaber, D., Wolff, M., & Daly, T. (2021). *Assessing the accuracy of elementary school test scores as predictors of students' high school outcomes* (CALDER Working Paper No. 235-0821-2). Center for the Analysis of Longitudinal Data in Education Research. https://caldercenter.org/sites/default/files/CALDER%20WP%20235-0520-2.pdf.

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

Guthrie, G. (2015). The formalistic education paradigm in Papua New Guinea. *Contemporary PNG Studies*, *22*, 33.

Hanushek, E., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, *46*(3), 607–668.

Hershbein, B., Harris, B., & Kearney, M. (2014). *Major decisions: Graduates' earnings growth and debt repayment*. Hamilton Project, Brookings Institution. http://www.hamiltonproject.org/assets/files/major_decisions_graduates_earnings_growth_debt_repayment.pdf.

Jacob, R., Hill, H., & Corey, D. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 379–407.

Jacobse, A. E., & Harskamp, E. G. (2011). *A meta-analysis of the effects of instructional interventions on students' mathematics achievement*. GION, Gronings Instituut voor Onderzoek van Onderwijs, Opvoeding en Ontwikkeling, Rijksuniversiteit Groningen.

Koon, S., Foorman, B., & Galloway, T. (2020). *Identifying North Carolina students at risk of scoring below proficient in reading at the end of grade 3* (REL 2020–030). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. https://files.eric.ed.gov/fulltext/ED607256.pdf.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review*, *10*(3), 179–214.

Levin, H., McEwan, P., Belfield, C., Bowden, B., & Shand, R. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*. Sage Publications.

Loeb, S., & Bassok, D. (2007). Early childhood and the achievement gap. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 517–534). Routledge.

Magnuson, K. A., & Duncan, G. J. (2006). The role of family socioeconomic resources in the black–white test score gap among young children. *Developmental Review*, *26*(4), 365–399.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749.

Menyhart, O., Weltz, B., & Győrffy, B. (2021). MultipleTesting. com: A tool for life science researchers for multiple hypothesis testing correction. *PLOS One*, *16*(6), e0245824.

Mokher, C. G., Leeds, D. M., & Harris, J. C. (2018). Adding it up: How the Florida College and Career Readiness Initiative impacted developmental education. *Educational Evaluation and Policy Analysis*, *40*(2), 219–242.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind?. *Journal of Learning Disabilities*, *44*(5), 472–488.

Murnane, R. J., Willett, J. B., Duhaldeborde, Y., & Tyler, J. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings. *Journal of Policy Analysis and Management*, *19*(4), 547–568.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2001). Are effects of small classes cumulative? Evidence from a Tennessee experiment. *Journal of Educational Research*, *94*, 336–345.

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the

high school exit exam. *Economics of Education Review*, *29*(2), 171–186.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2011). *How performance information affects human-capital investment decisions: The impact of test-score labels on educational outcomes* (NBER Working Paper No. 17120). National Bureau of Economic Research.

Pellegrini, M., Lake, C., Inns, A., & Slavin, R. E. (2018, October). *Effective programs in elementary mathematics: A best-evidence synthesis*. Annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, *37*(1), 239–253.

Petscher, Y., Fien, H., Stanley, C., Gearin, B., Gaab, N., Fletcher, J. M., & Johnson, E. (2019). *Screening for dyslexia*. U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy.

Randel, B., Apthorp, H., Beesley, D., Clark, F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *The Journal of Educational Research*, *109*(5), 491–502.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy*, *5*(2), 251–281.

Savelsbergh, E. R., Prins, G. T., Rietbergen, C., Fechner, S., Vaessen, B. E., Draijer, J. M., & Bakker, A. (2016). Effects of innovative science and mathematics teaching on student attitudes and achievement: A meta-analytic study. *Educational Research Review*, *19*, 158–172.

Schoen, R. C., LaVenia, M., & Tazar, A. M. (2018, March). *Effects of a two-year Cognitively Guided Instruction professional development program on first- and second-grade student achievement in mathematics*. Paper presented at the annual meeting of the Society for Research in Effective Education, Washington, DC.

Slavin, R. E., & Karweit, N. L. (1985). Effects of whole-class, ability grouped and individualized instruction on mathematics achievement. *American Educational Research Journal*, *22*(3), 351–367.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*(3), 427–515.

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1–26.

Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating math recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement. *American Educational Research Journal*, *50*(2), 397–428.

Stevens, R. J., & Slavin, R. E. (1995). The cooperative elementary school: Effects on students' achievement, attitudes, and social relations. *American Educational Research Journal*, *32*(2), 321–351.

Stipek, D. (2002). At what age should children enter kindergarten? A question for policy makers and parents. *Society for Research in Child Development Policy Report*, *16*, 13–16.

Styers, M., & Baird-Wilkerson, S. (2011). *A final report for the evaluation of Pearson's focusMATH Program*. Magnolia Consulting.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199.

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162–181.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of tier 2 type reading interventions in grades K-3. *Educational Psychology Review*, *28*(3), 551–576.

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2017). Does early mathematics intervention change the processes underlying children's learning?. *Journal of Research on Educational Effectiveness*, *10*(1), 96–115.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352–360.

Weis, R., Osborne, K. J., & Dean, E. L. (2015). Effectiveness of a universal, interdependent group contingency program on children's academic achievement: A countywide evaluation. *Journal of Applied School Psychology*, *31*(3), 199–218.

Xu, Z., Özek, U., & Hansen, M. (2015). Teacher performance trajectories in high-and lower-poverty

schools. *Educational Evaluation and Policy Analysis*, *37*(4), 458–477.

Yeh, S. S. (2010). The cost effectiveness of 22 approaches for raising student achievement. *Journal of Education Finance*, *36*(1), 38–75.

Authors

ZEYU XU, PhD, is a managing economist at the American Institutes for Research. His research focuses on college readiness, high school-to-college transition, and teacher labor market policies.

UMUT ÖZEK, PhD, is a senior economist at RAND. His current research focuses on immigrant students and English learners, implementation and consequences of educational accountability, K–12 remediation policies and their effects, and the design and effects of school choice programs.

JESSE LEVIN, PhD, is a principal research economist at the American Institutes for Research. His research focuses on the economic evaluation of educational programs and investigating the adequacy and equity of school funding mechanisms.

DONG HOON LEE, MPP, is a researcher at the American Institutes for Research. His research focuses on inequalities in schools among immigrants and English learners, as well as policy solutions and their effects on these population groups.

Manuscript received June 23, 2023
First revision received March 6, 2024
Second revision received July 8, 2024
Third revision received August 27, 2024
Accepted October 25, 2024