Research Article

Educational Evaluation and Policy Analysis
Month 202X, Vol. XX, No. X, pp. 1–28
DOI: 10.3102/01623737251400377

Impacts of Academic Recovery Interventions on Student Achievement in 2022 to 2023

Maria V. Carbonari
Harvard University

Michael DeArmond
American Institutes for Research, CALDER

Daniel Dewey
Harvard University

Elise Dizon-Ross
Dan Goldhaber
American Institutes for Research, CALDER

Thomas J. Kane
Harvard University

Anna McDonald
American Institutes for Research, CALDER

Andrew McEachin
NWEA

Emily Morton
American Institutes for Research, CALDER

Atsuko Muroga
Harvard University

Alejandra Salazar
American Institutes for Research, CALDER

Douglas O. Staiger
Dartmouth College

The COVID-19 pandemic caused large, persistent declines in student achievement. This paper examines 2022–2023 recovery efforts across eight districts, including tutoring, small-group instruction, after-school, extended year, double-dose, digital learning, and expert teacher interventions. The study includes 12 math programs and 15 literacy programs (three covered both subjects). Most served fewer students and provided lower dosage than planned. Using value-added models, we find positive, robust effects in just five subject-intervention pairs: one subject-specific tutoring program (+0.22 SD math, +0.22 SD reading), a small-group reading intervention (+0.33 SD), and assignment to "expert" teachers (+0.06 SD math, +0.11 SD reading). Results highlight the promise of intensive interventions while underscoring the difficulty of scaling them to meet recovery needs.

Keywords: *achievement, COVID-19, academic intervention, tutoring, K–12 education policy*

The COVID-19 pandemic had a significant negative impact on student achievement, with nationwide average declines comparable to those observed after Hurricane Katrina (Sacerdote, 2012). Pandemic-related disruptions to public schooling and other social services disproportionately affected students from historically marginalized groups. Achievement gaps widened, arguably undoing nearly 2 decades of progress toward educational equity in the United States (U.S. Department of Education, 2021, 2022). As of the spring of 2023, 3 years after the initial pandemic-related school closures, average achievement levels remain well below pre-pandemic norms, especially for students of color and students in high-poverty districts (Curriculum Associates, 2023; Fahle et al., 2023, 2024; Goldhaber et al., 2023; Lewis & Kuhfeld, 2022, 2023, 2024).

Supported by $190 billion of Elementary and Secondary School Emergency Relief (ESSER) funds, school districts responded to pandemic losses with a range of academic interventions, including tutoring, push-in or pull-out small-group instruction, before- and after-school programs, summer learning programs, and extended school days and years (Diliberti & Schwartz, 2022). Early evidence on pandemic-recovery initiatives showed the challenge of quickly ramping up programs for large numbers of students. For example, studies suggest that academic interventions in the pandemic's initial years reached fewer students than planned and provided the average participant with fewer hours of support than intended. School districts also confronted a complex mix of implementation issues, including scheduling problems, staffing shortages and absenteeism, and inadequate central office capacity (Carbonari et al., 2024; Makori et al., 2024). Early evidence from commonly used interim assessments suggests that districts' academic interventions did not substantially improve the pace of student achievement growth during the 2021–2022 school year (Barry & Sass, 2022; Callen et al., 2023; Carbonari et al., 2024; Robinson et al., 2022).

By the summer of 2022, however, signs of improvement started to emerge. Analyzing the academic progress of students who attended summer school in eight school districts, Callen et al. (2023) found a positive impact for summer school in 2022 on math test achievement (+0.03 *SD*) but not in reading. Using data from state tests, Fahle et al. (2024) found evidence of recovery from spring 2022 to spring 2023 in math and reading across 29 states.[1] Using the same state test score data, recent analyses suggest that the districts that received more ESSER funds exhibited faster growth between spring 2022 and spring 2023, contributing to their recovery (Dewey et al., 2024; Goldhaber & Falken, 2024). Despite this progress, spring 2023 test scores on interim assessments remained far below what we would have expected based on pre-pandemic achievement levels, with historically marginalized students falling the furthest behind their pre-pandemic performance levels (Curriculum Associates, 2023; Lewis & Kuhfeld, 2023).

The stakes surrounding students' academic recovery remain high. Hanushek and Strauss (2024), for example, estimate that unremedied declines in students' test scores could translate to reductions in average lifetime earnings of 2% to 9% for students and a 3.5% decrease in economic growth, totaling $31 trillion. Doty et al. (2022) forecast smaller (but still large) impacts when

limited to individual students' lifetime earnings ($900 billion). In either case, the fallout is likely to be even more severe for students of color and economically disadvantaged students. Four years after COVID disrupted schools, learning acceleration and academic recovery continue to be crucial issues for the U.S. economy and social equality (The White House, 2024).

This paper extends our prior analyses of COVID recovery (Callen et al., 2023; Carbonari et al., 2024) by examining recovery efforts in eight school districts during the 2022–2023 school year. The eight districts are part of the Road to Recovery (R2R) research project, a partnership that began in 2021 between the districts and researchers at NWEA, Harvard University, and the American Institutes for Research.

We examine interventions that were designed to deliver supplemental instructional time to students, including tutoring programs, after-school programs, digital learning programs, extended school years, double-dose classes, and push-in and pull-out instruction for small groups of students (i.e., "interventionists"). In one district, we also examine a less common intervention that did not provide students with additional time: assigning struggling students to "expert" teachers with high evaluation scores (based in part on prior test-based value-added). In all cases, we focus on targeted academic recovery interventions for subsets of students, rather than universal programs. Focusing on targeted interventions allows us to analyze intervention effects by comparing participating students to non-participating students (i.e., we do not study district-wide interventions affecting all students in a grade(s), such as math coaching for elementary teachers or a new literacy curriculum).

Across the eight school districts, we examine 12 interventions that provide math instruction and 15 interventions that provide literacy instruction[2] across grades K–8.[3] We categorize the interventions into three groups: (a) tutoring and small group instruction, (b) other supplemental instruction time programs, and (c) the "expert teacher" program. Of these interventions, three provide math and literacy instruction to all participants (e.g., after school, extended school year), and nine enroll students specifically into math and/or literacy support (e.g., tutoring, digital learning). Three additional interventions provide instruction only in reading. While two districts in our sample had similar expert teacher programs that assigned students to teachers designated as expert teachers in math or reading based on observations, student growth, and National Board certification, we received data on the program from just one district. We mostly rely on observational methods to evaluate the effects of the programs on students' math and reading achievement, comparing treated and untreated students' scores after controlling for students' characteristics and prior test scores. In a few instances, where the program design and data allow, we use a regression discontinuity (RD) design.

Consistent with prior studies of pandemic-era interventions, we find most of the interventions served fewer students than intended and delivered fewer hours than planned (Barry & Sass, 2022; Carbonari et al., 2024; Makori et al., 2024; Robinson et al., 2022) and/or did not positively impact student achievement (Pollard et al., 2024). Across all tutoring, small group instruction, and other supplemental instruction time interventions in the present study, we estimate positive effects of just one tutoring program for math and two tutoring programs for reading. Relative to the other programs in our sample, these served far fewer students (~1%–2% of students in eligible grades) and provided students with more instruction time over the course of the year (>30 hours). We also estimate a significant negative effect of one tutoring program in math. We also find that having an expert teacher improves achievement significantly more than having a non-expert teacher. Collectively, our findings highlight the promise of intensive academic interventions while underscoring the challenges school districts face implementing them on a scale commensurate with the pandemic's impact.

Background

Pre-Pandemic Evidence on Effective Academic Interventions

The pre-pandemic literature on academic interventions highlights several strategies that could help students' academic recovery. High-impact tutoring programs (Nickow et al., 2024), summer learning programs (Kim & Quinn, 2013; Lynch et al., 2023; McCombs et al., 2014), and

double-dose math courses (Nomi & Allensworth, 2013) all have strong pre-pandemic evidence improving student achievement.

Other popular interventions from the pre-pandemic period have a more mixed evidence base. These include: after-school programs (McCombs et al., 2017), computer-assisted learning (CAL) programs (Bettinger et al., 2023; Escueta et al., 2017), extended school days or years (Checkoway et al., 2013; Kraft, 2015; Kraft & Novicoff, 2024), double-dose courses in literacy (Arthur & Davis, 2016; Nomi, 2015; Özek, 2021), and grade retention (Jacob & Lefgren, 2009; Opper & Özek, 2024; Özek & Mariano, 2023). In these cases, some studies report significant achievement gains, while others find null effects or even unintended negative consequences.[4]

Previous literature also provides some useful guidance on the design of effective interventions. Recent tutoring meta-analyses (Kraft, Schueler, & Falken, 2024; Nickow et al., 2024), for example, note more effective tutoring programs tend to use teachers or paraprofessionals as tutors, serve students in earlier grades, deliver tutoring during the school day, and occur at least 3 days per week (at least 15 hours in total).[5] However, few of the pre-pandemic tutoring studies evaluated programs that were delivered on the scale that would arguably be needed for COVID recovery, raising questions about whether high-fidelity implementation is feasible (or necessary) for delivering promising programs at scale. Indeed, Kraft, Schueler, and Falken's (2024) meta-analysis of pre-pandemic tutoring randomized control trials (RCTs) finds just 11% of the 265 programs in their study served over 400 students, and these larger programs had impacts approximately two-thirds the size of programs serving fewer students. Similarly, just 17% of the 89 tutoring programs in Nickow et al.'s (2024) meta-analysis had a sample size greater than 400 students—only some of whom received the treatment—and effect sizes were smaller for larger programs. This negative association between program size and effectiveness is also observed more broadly across education research (Kraft, 2020). Reduced oversight and increased variation in implementation when programs serve more students may help explain this pattern (Hill & Erickson, 2019). Thus, while the pre-pandemic literature highlights promising programs, it does not provide a clear road map on how to scale them up for COVID recovery.

Implementing and Scaling Pandemic-Era Academic Recovery Interventions

During the second and third years of ESSER (2021–2022 and 2022–2023), school districts struggled to implement and scale interventions effectively. They grappled with scheduling conflicts, staffing shortages, limited staff capacity, insufficient central office management, and inadequate data systems (Carbonari et al., 2024; Makori et al., 2024). Schools adapted interventions to fit their resources and needs, sometimes targeting the wrong students, replacing core instructional time, or deviating from central-office intentions or evidence-based practices (Carbonari et al., 2024). The 2021–2022 school year was particularly challenging, with new COVID variants and political polarization further complicating implementation efforts.

Some adaptations may have been necessary to reach more students, but they also may have compromised efficacy. For example, a district with a limited number of highly qualified tutors trying to scale up their tutoring program may face a trade-off between reducing tutoring hours per student or increasing tutoring group sizes, both of which may reduce program effectiveness. Kraft, Schueler, and Falken (2024) identify a bundle of program features that protects against the attenuation of tutoring impacts as programs expand: in-person instruction, delivered during school hours and on school premises, with a maximum three-to-one student-to-tutor ratio, meeting at least three times per week and at least 15 hours in total, and using a provided curriculum. They find suggestive evidence that the bundle of features, with in-person programming and a minimum of 15 hours of total instruction being most important, is key to guarding against effect attenuation for programs serving 400 to 999 students.

The limited existing research on pandemic-era academy recovery interventions, which has focused on tutoring programs, also suggests potential tradeoffs between program size, dosage, and effectiveness (see Supplemental Appendix Table A1 in the online version of the journal). In our prior study of 2021–2022

interventions in four of the R2R districts, the smallest tutoring program in the sample still served nearly 400 students, and most programs served over 1,000 students (Carbonari et al., 2024). Only one program delivered over 15 hours of dosage, and it was one of the smallest (497 students tutored) programs and had the largest positive effect size (0.048 *SD*), though it was not statistically significant. None of the other tutoring programs in the study had statistically significant, positive impacts on student achievement.

A study of pandemic-era tutoring in Metro Nashville Public Schools further highlights the potential tradeoffs between scale, dosage, and impact (Kraft, Edwards, & Cannata, 2024). During the 2021–2022 school year, 2,612 students received an average of ~12 hours of tutoring in either the first or second semester, and an RCT of the fall 2021 impacts estimates null effects on test scores. In 2022–2023, however, the district increased both the scale (7,296 students) and dosage (~18 hours) of the program. Using a quasi-experimental difference-in-differences approach to estimate average effects of the program over both years, Kraft, Edwards, and Cannata (2024) found small, positive impacts on reading achievement (0.04–0.09 *SD*) and null effects on math. While the positive effect in reading is promising, it remains well below the average effect of pre-pandemic programs serving 1,000 students or more (0.16 *SD*). The authors hypothesize that some of the reduced impact may be due to flexible program design features (i.e., sessions could be scheduled during or after school, could be online or in-person), limited treatment-control contrast (i.e., control group students were also receiving small group supports), and heterogeneous effects across students (i.e., larger impacts for higher-performing students), in addition to relatively low dosage.

Ready et al. (2024) provide another example of these tradeoffs, finding small impacts for a large-scale and low-dosage virtual reading tutoring program serving students in grades 1 to 4. Their RCT estimates a 0.05 *SD* increase in reading achievement on NWEA's Measures of Academic Progress (MAP) Growth assessment for 959 treated students (study sample = 1,777) across six schools. The program offers two to three sessions per week for 30 minutes per session over 10 weeks (10–15 hours of treatment in total). Sessions occur during a daily class period used for intervention and acceleration (i.e., "Learning Lab"). However, only about 20% of treated students completed the recommended dosage of at least 10 hours, with an average of just 6.5 hours received. Higher-performing students spent significantly more time using the program, and researchers also found a positive correlation between dosage and treatment effects.

In contrast to these programs with low dosage or participation, two pandemic-era programs achieved high dosage, high participation, and substantial impacts. Interestingly, both programs leveraged technology to help them deliver programming at scale. Cortes et al. (2025), for example, conducted an RCT of the Chapter One literacy tutoring program, which served 420 students (study sample = 818) across 49 kindergarten classrooms and found that the program improves students' reading achievement by 0.11 *SD*. The technology-driven program costs $450/student and involved part-time tutors "pushing-in" to the classroom to provide short bursts (5–10 minutes) of instruction to individual students up to five times per week for the duration of the school year, as well as 10 minutes using Chapter One's software every day, totaling approximately 27 hours per year on average.[6] An RCT of an even larger-scale and higher dosage tutoring program found 0.23 *SD* gains in math scores for 2,060 ninth-grade students (study sample = 3,846) across two districts who took part in daily tutoring for 50 minutes per day (~150 hours per year; Bhatt et al., 2024). The program had pairs of students alternated between receiving in-person tutoring and CAL each day, reducing costs (~$2,200 per student vs. ~$3,500 per student without CAL) and increasing its scalability.

Despite some promising examples, recent literature underscores the initial challenges school systems faced implementing pandemic-era interventions, especially at scale. However, it provides limited evidence on how targeted recovery efforts fared in 2022 to 2023, when school districts largely returned to normal operations and were presumably better positioned to fully implement their interventions. We also hypothesize that 2022–2023 programs would reach more students, be implemented with greater fidelity, and be more effective at improving student

achievement than their 2021–2022 counterparts, as districts would have had the opportunity to adapt their programming in response to the challenges of the previous year. This study provides insights into the tradeoffs between intervention type, scale, dosage, and effectiveness across eight large school districts implementing different academic recovery interventions in 2022 to 2023. Given the enduring impacts of the pandemic on student achievement and the end of ESSER funding in September 2024, evidence on the tradeoffs between different pandemic-era interventions is critical for informing districts' and states' ongoing, urgent recovery efforts.

Methods

Sample

We examine academic recovery in the 2022–2023 school year for a sample of K–8 students from eight districts—Alexandria City Public Schools (VA), Dallas Independent School District (TX), Guilford County Schools (NC), Portland Public Schools (OR), Richardson Independent School District (TX), Suffern Central School District (NY), Syracuse City School District (NY), and Tulsa Public Schools (OK)—participating in the R2R project.[7] We veil districts' names when reporting district-specific demographics or results to protect their anonymity. We also aim to describe program designs with sufficient detail without inadvertently disclosing district identities. As displayed in Table 1, the eight districts collectively enroll over 360,000 K–12 students across six states. They serve higher proportions of Black and Hispanic students, and students eligible for free or reduced-price lunch than the national average.

Three of the eight districts have publicly available COVID-recovery achievement data published on the Education Recovery Scorecard website for 2023 (Reardon et al., 2024). These data allow for national comparisons of the districts' academic recovery by linking state test proficiency scores to the NAEP in 2022. In Table 2, we show the extent to which the study districts' test scores in math and reading had recovered to pre-pandemic (i.e., spring 2019) levels as of spring 2023. Whereas Guilford County Schools have a slightly smaller remaining gap (−0.06 *SD*) than the average district (−0.08 *SD*) in math and a moderately larger remaining gap (−0.12 *SD*) than the average district (−0.06 *SD*) in English Language Arts (ELA), Alexandria and Tulsa have substantially larger remaining gaps, ranging from −0.28 to −0.39 *SD* across math and ELA. Based on average yearly pre-pandemic gains on interim assessments across grades 3 to 8, these larger declines around −0.3 to −0.4 *SD* are roughly equivalent to 80% to 110% of the gains students make in a typical year (Kuhfeld et al., 2023).

Data

This study uses NWEA's MAP Growth test scores, district-provided student-level data on demographics and eligibility for and participation in academic recovery interventions. We also use information collected through interviews of district leaders about the design of interventions to characterize the interventions we analyze.

Test Scores. Our math and reading achievement outcomes are student test scores on NWEA MAP Growth assessments in grades K–8. The MAP Growth assessment is an interim assessment administered to students three times each year (fall, winter, and spring), which allows us to observe changes in student achievement within the school year. The timing of the tests is helpful for evaluating interventions that were administered to students for less than a full school year, that is, in the fall semester, spring semester, or during the summer. It is also a computer adaptive assessment, responding to a student's performance throughout the test event. Adaptability increases test score precision, especially at the tails of the distribution. This feature is particularly important in the context of the pandemic, when many more students are performing below grade-level.

We use the NWEA 2020 MAP Growth norms (Thum & Kuhfeld, 2020) to standardize the MAP scores by subject and grade.[8] NWEA calculated these norms using MAP scores from a nationally representative sample of students from three pre-pandemic school years (i.e., 2016–2017, 2017–2018, and 2018–2019). We compare students' pandemic-affected test scores to the national pre-pandemic test distribution. The NWEA database also includes information on students' race/ethnicity and gender, and school-level NCES

TABLE 1

Sample Demographics

Demographic	Study districts	Nationwide NWEA districts	U.S. public schools
Average district enrollment	45,825	—	2,766
Average school enrollment	583	484	514
FRPL eligible (%)	70	54	50
Race (%)			
Asian	4	4	5
Hispanic	39	21	28
Black	26	15	15
White	25	53	44
School locale (%)			
City	86	29	30
Suburb	8	32	39
Town	0	11	11
Rural	5	29	20

Note. Data for the national sample and study district sample are from the CCD collected by the National Center for Education Statistics during the 2022–2023 school year. Statistics for the Nationwide NWEA sample are based on data from the 2019 to 2020 CCD data collection, as reported in Isaacs et al. (2023). CCD = Common Core of Data; FRPL = free or reduced priced lunch.

identifiers that we link to school-level enrollment and demographic data from the 2020–2021 Common Core of Data.

District-Provided Student-Level Data. Each district provided student-level data on demographics, intervention eligibility, state test scores, and intervention participation. These data allowed us to identify which students participated in each intervention, to report the hours of instruction students attended or received in each intervention (by subject when possible), and to estimate the impacts of each intervention on students' spring 2023 MAP scores.

Intervention Design Interviews. We collected detailed programmatic information on interventions from interviews with central office intervention leaders, district-provided documents, and information available on districts' public websites. We asked districts to identify academic recovery interventions that met all the following criteria: (a) interventions were new or expanded since the pandemic, (b) interventions were supported by ESSER funds, and (c) interventions provided targeted students with additional learning time beyond what was offered during standard instruction.

Districts also shared contact information for the district-level leader(s) of each of these interventions, with whom we conducted virtual, semi-structured interviews in spring 2023 that lasted between 30 and 90 minutes. The interviews included questions about program content, program intensity, delivery mode, program providers, and student eligibility criteria.[9] Across the eight districts, we identified seven categories of academic recovery interventions: (a) tutoring programs, (b) small-group push-in and pull-out interventions, (c) after-school programs, (d) extended school years, (e) double-dose classes, (f) digital learning programs, and (g) assignment to an expert teacher.[10]

The interventions implemented in each of the eight districts are displayed in Table 3. In some cases, students could participate in multiple interventions at once (e.g., extended year and tutoring). We do not analyze the impacts of District C's extended year calendar because the extra school days added to the calendar at a subset of schools did not occur between fall and spring MAP Growth testing periods.[11]

TABLE 2

Estimated Achievement Loss and Recovery From Spring 2019 to 2023, Grades 3 to 8

Subject	District	Spring 2019 (*SD*s)	Spring 2022 (*SD*s)	Spring 2023 (*SD*s)	Change from S19 to S22 (*SD*s)	Change from S22 to S23 (*SD*s)	Change from S19 to S23 (*SD*s)
Panel A: Math	Alexandria	−0.10	−0.50	−0.48	−0.40	0.02	−0.38
	Dallas	−0.04	−0.22	—	−0.18	—	—
	Guilford	−0.11	−0.21	−0.17	−0.11	0.04	−0.06
	Portland	—	—	—	—	—	—
	Richardson	0.25	0.05	—	−0.20	—	—
	Suffern Central	—	—	—	—	—	—
	Syracuse	—	—	—	—	—	—
	Tulsa	−0.67	−1.08	−1.06	−0.41	0.01	−0.39
	Study district average	−0.12	−0.32	−0.57	−0.21	0.03	−0.28
	National district average	0.05	−0.08	−0.03	−0.13	0.05	−0.08
Panel B: ELA	Alexandria	−0.14	−0.37	−0.42	−0.22	−0.06	−0.28
	Dallas	−0.21	−0.38	—	−0.16	—	—
	Guilford	−0.03	−0.16	−0.14	−0.13	0.02	−0.12
	Portland	—	—	—	—	—	—
	Richardson	0.00	−0.11	—	−0.11	—	—
	Suffern Central	—	—	—	—	—	—
	Syracuse	—	—	—	—	—	—
	Tulsa	−0.60	−0.96	−0.94	−0.36	0.01	−0.34
	Study district average	−0.16	−0.32	−0.50	−0.16	−0.01	−0.25
	National district average	0.06	−0.03	0.04	−0.09	0.03	−0.06

Note. All estimates are from the Stanford Education Data Archive (Version SEDA 2023 2.0; Reardon et al., 2024) and are scaled such that a 0 in this metric is equal to the average of the national NAEP average (in grade 5.5) in spring 2019, and 1 unit in this metric is equal to 1 student-level *SD*. Estimates in this scale are comparable across the whole country, and over time, but they are not comparable across subjects. "—" indicates achievement data for the relevant district, subject, and time point(s) were not available in the SEDA dataset. *SD* = standard deviation.

TABLE 3

Program Usage Across Sample Districts

District	Tutoring	Small group	After-school	Extended calendar	Double-dose	Digital learning
Panel A: Math interventions						
District A	X			X		
District B	X		X			
District C	X X			X		
District D						X
District E		X				
District F	X					
District G		X				
District H			X			
Panel B: Reading interventions						
District A	X	X		X		
District B	X		X			
District C	X X			X		
District D						X
District E		X				
District F	X				X	
District G	X	X				
District H			X			

Note. We do not disclose the district that implemented the expert teachers intervention to preserve district anonymity.

We provide detailed information about the design characteristics (e.g., targeting criteria, delivery modality, provider type) of the (a) tutoring and small group instruction programs and (b) after-school, extended school year, double-dose classes, digital learning, and expert teacher programs respectively in Supplemental Appendix Tables A2 and A3 (available in the online version of this article).

The interventions vary in their designs both across and within program types and the number of students served. For tutoring and small group instruction interventions, most programs used test scores to target students in some capacity. Most commonly, programs intended to serve all students who scored below a certain threshold, but some programs targeted students performing with a particular range of scores. Most districts opted to deliver their tutoring and small group interventions in-person (as opposed to virtually), during school hours, and with a max provider-to-student ratio of 1:6. Small group instruction programs employed certified district staff at higher rates than tutoring programs, while the latter relied on a variety of providers, including district staff, college students, community members, and even high school students. These interventions varied widely in their intended dosage, ranging from ~9 to ~134 hours.

Extended school year, after-school, digital learning, double-dose, and expert teachers used larger groups or a classroom setting, with provider-to-student ratios above 1:10 (with the exception of District D's digital learning intervention; see Supplemental Appendix Table A3 in the online version of the journal). These interventions also varied widely in terms of their total intended dosage, ranging from ~9 to ~124 hours over the course of the year.

Empirical Approach

Value-Added Models. We estimate the impact of all interventions using value-added models (VAMs) that control for observable baseline student characteristics and test scores (although, as discussed below, we also estimate impacts using an RD design in a subset of cases). VAMs have often

been used to estimate the impacts of schools on student outcomes (McEachin & Atteberry, 2017) as well as the impacts of interventions and policies on student achievement (Barry & Sass, 2022). We specify the following equation:

$$\begin{aligned} MAP_{igj,sub}^{Sp2023} = {} & \beta_0 + \beta_1 Int_{i,sub} + \beta_2 Int_{i,sub}^{other} \\ & + \beta_3 Int_{i,\widetilde{sub}} + \delta MAP_{i,sub}^{Fa2022} Grade_i \\ & + \tau X_i Grade_i + \psi_{gj} + \varepsilon_{igj} \end{aligned}$$

where $MAP_{igj,sub}^{Sp2023}$ is the MAP Growth score for student i in grade g in school j in subject sub in spring of 2023. We standardize all MAP Growth scores at the subject and grade level using NWEA MAP Growth pre-pandemic norms, so that the units are in standard deviations of the national distribution of student MAP performance prior to the COVID outbreak.[12] $Int_{i,sub}$ is a binary indicator of intervention participation for the intervention in question. $Int_{i,sub}^{other}$ and $Int_{i,\widetilde{sub}}$ are vectors of binary indicators of intervention participation for all other available interventions in subject sub and in other subject $\widetilde{sub}$. We use the other subject scores as a form of placebo test to assess the potential for selection bias, as we discuss more below. We include controls for participation in all available interventions in order to estimate the effect of each program *individually*, as students frequently participate in multiple interventions throughout the year. The coefficient of interest is $\hat{\beta}_1$, the estimated average treatment effect of the intervention in question. To account for the increased risk of Type I error from estimating this specification across 22 subject-intervention pairs, we assess the statistical significance of $\hat{\beta}_1$ using the Bonferroni correction, which adjusts the significance threshold by dividing alpha by the number of comparisons (i.e., $0.05/22 = 0.0023$). All statistically significant estimates discussed in the text ($p < .01$ for each of these estimates) are also significant under this adjusted threshold ($p < .0023$).

To control for baseline achievement, we include in our regressions $MAP_{i,sub}^{Fa2022}$, a cubic polynomial function of student i's norm-standardized MAP Growth score at the start of the school year, interacted with student i's grade level. Our analytic sample for each intervention is restricted to students enrolled in grades that were eligible for the intervention with non-missing MAP Growth scores in fall 2022 and spring 2023. The vector X_i includes student i's available baseline demographics (i.e., indicators for student race/ethnicity, gender, Individualized Education Program status, English language learner status, 504 plan status, and economic disadvantage status), indicators for the calendar week they took MAP Growth tests in fall 2022 and spring 2023, linear functions of prior MAP Growth scores from winter 2022 and spring 2022 in subject sub, and a linear function of MAP Growth scores from fall 2022 in the opposite subject $\widetilde{sub}$.[13] Because we allow for missingness in these earlier and opposite subject test scores, we interact all test scores with indicators for missingness. We additionally interact all elements of vector X_i with grade level. Finally, Ψ_{gj} denotes school-by-grade fixed effects and ε_{igj} represents idiosyncratic error. We estimate a linear model and calculate standard errors while clustering at the school-by-grade level (Abadie et al., 2023).[14]

The inclusion of baseline and prior achievement scores is critical in our estimation strategy, given the selection of students into interventions. This selection is illustrated for the tutoring and small group interventions and other supplemental time interventions respectively in Supplemental Appendix Figures A1 and A2 (available in the online version of this article). The figures show the average difference in MAP Growth scores between treated and untreated students in each intervention, controlling for student demographics, in prior terms leading up to the start of the 2022–2023 school year. With a few exceptions, the mean difference is consistently negative, as would be expected for interventions aimed at improving academic achievement for lower performing students.[15] The mean difference is not consistently negative for the expert teacher program, which did not target students by performance, and Districts B and F's tutoring programs, which targeted low-performing students who scored within a particular range but were not available to the lowest performing students in the districts.

Even after controlling for multiple prior test scores, however, there remains the possibility that our models omit other relevant variables related to both treatment assignment and student outcomes, biasing our estimates of the interventions' treatment effects. This is an inherent

weakness of all selection-on-observables research designs. In this case, for example, we typically do not have access to data on program referrals by teachers, which in many districts were informed not only by observable academic measures but also by teachers' subjective judgments of students' needs. Because we cannot say with certainty that our VAMs capture all relevant variables, we do not interpret the results from these models as causal estimates, but rather as comparisons of outcomes for treated and untreated students who are as observably similar as possible. Recognizing this limitation, we do, however, use the terms "effects" and "impacts" throughout the paper for simplicity in writing.

We also conduct multiple tests to assess or limit the role that selection might play in our treatment estimates. The first is that we conduct placebo tests estimating the effects of participating in a subject-specific intervention on test scores in the other subject. While we cannot definitively rule out cross-subject impacts of interventions, this placebo test provides us with a measure of the potential for selection bias. Specifically, we estimate Equation 1 and replace the outcome variable (i.e., math or reading) with MAP Growth test scores in the other subject (i.e., reading or math). The point estimates of these tests can be interpreted as estimates of selection bias under the following two assumptions: (a) that participating in a subject-specific intervention does not affect students' scores in the other subject and (b) that students' gains in the intervention subject would have trended similarly to their gains in the other subject if they had not participated in the intervention.

Secondly, for interventions where we detect positive and statistically significant impacts for participants, we calculate bounds following Oster (2019) to assess whether our estimates are robust to adjusting for potential omitted variable bias. These bounds assume a maximum *R*-squared of 1 in the hypothetical scenario where all relevant covariates were included in the model, as well as assume that the influence of unobservable (i.e., omitted) variables is no greater than the influence of observable variables included as controls in the model. We describe this sensitivity test in greater detail and present our bounded estimates in Supplemental Appendix B (available in the online version of this article).

Thirdly, out of concern that differential attrition could bias our treatment estimates (e.g., if struggling students are more likely to be assigned to treatment and also more likely to be dropped from the sample for missing MAP Growth scores) we assess the missingness of both baseline and outcome test scores, by treatment status, for each of the interventions we evaluate. Comparing these rates (as shown in Supplemental Appendix Table A7 in the online version of the journal), we see that missingness rates are generally higher among untreated students than among treated students. If we assume that lower performing students are more likely to be missing data—either because they did not take the assessment, took an alternative version of the assessment, and/or dropped out of the district—this observation would suggest that our estimated treatment effects are conservative and could potentially be biased downward. That said, the differences in missingness rates between treated and untreated students are generally small, and it is unlikely that missingness is uniformly more prevalent among lower-performing students.

Finally, in addition to these robustness checks, we use meta-analysis methods to understand the average effect of interventions in each subject across districts. Specifically, we use a random effects model with restricted maximum likelihood estimation to generate the overall estimates (DerSimonian & Laird 1986; Hedges, 1983; Raudenbush, 2009). This approach assumes that our treatment effect estimates are unbiased, which, as we discussed above, we recognize may not be the case for all interventions.

Regression Discontinuity. For certain interventions, we are able to estimate treatment effects using a fuzzy RD design. Researchers can apply the RD method when there is a clearly defined cutoff for intervention eligibility (e.g., test scores, date of birth). RD designs estimate causal effects of interventions by comparing the outcomes of interest for observations just above and below the cutoff.

Two districts (Districts A and D) identified students for interventions using their standardized state test scores, with a cutoff point demarcating eligibility.[16] Districts identified students whose scores fell below this cutoff as eligible for

the intervention. Students with scores above the cutoff were ineligible to participate. We use this "jump" in the probability of receiving the intervention at the cutoff to estimate the intervention effects. Because districts did not strictly adhere to this cutoff, we employ a fuzzy RD design to account for the presence of non-compliers (i.e., ineligible students who participated in the intervention and a small number of eligible students who did not participate). A fuzzy RD design adjusts the treatment effect near the cutoff for the level of non-compliance, which is the actual difference in participation at the cut-off (Hahn et al., 2001). The resulting impact estimate applies to those students who complied with their treatment assignment.

For subject *sub*, we specified the following two-stage least squares model:

$$\begin{aligned}\text{Any}_{\text{sub},i} &= \alpha_0 + \alpha_1 \text{Elig}_{\text{sub},i} + \alpha_2 \text{Score}_{\text{sub},i} \\ &\quad + \alpha_3 \left(\text{Score}_{\text{sub},i} * \text{Elig}_{\text{sub},i}\right) + \gamma_i + e_i\end{aligned}$$

$$\begin{aligned}\text{MAP}^{\text{Sp2023}}_{\text{sub},i} &= \beta_0 + \beta_1 \widehat{\text{Any}}_{\text{sub},i} + \beta_2 \text{Score}_{\text{sub},i} \\ &\quad + \beta_3 \left(\text{Score}_{\text{sub},i} * \text{Elig}_{\text{sub},i}\right) + \gamma_i + u_i\end{aligned}$$

where $\text{Any}_{\text{sub},i}$ is an indicator variable that takes the value 1 if student *i* received any treatment in subject sub during school year 2022–2023. $\text{Elig}_{\text{sub},i}$ is a binary eligibility indicator that equals to 1 if student *i* scores below the test score cutoff in subject sub, making the student eligible for the intervention. $\text{Score}_{\text{sub},i}$ is the "running variable," student *i*'s score of the test in subject *sub* that determined the eligibility for the intervention (standardized within the district so that the unit is in the standard deviation unit, and centered on the cutoff). γ_i represents the grade-by-language (i.e., the language of the standardized test that determined the intervention eligibility) fixed effects and e_i, the idiosyncratic error.

The second stage model estimates the outcome, spring 2023 MAP, denoted as $\text{MAP}^{\text{Sp2023}}_{\text{sub},i}$ and uses the estimated $\widehat{\text{Any}}_{\text{sub},i,t}$ from the first stage. The parameter of interest β_1 represents the local average treatment effect of being assigned to the intervention. We used triangular kernel weighting and estimated local nonparametric regression clustering standard errors at the school-by-grade level. To examine the sensitivity of our estimates by the bandwidths, we used 0.50, 0.75, and 1.00 *SD* of the running variable.

Results

Participation and Dosage

Table 4 summarizes the participation rates and dosage received for all interventions examined in this study. Column 2 shows the percentage of students in eligible grades in districts targeted for an intervention. In some cases, districts did not provide clear guidelines for identifying students for interventions. In other cases, districts did not have the relevant data for eligibility decisions. In both cases, we left these cells blank. Column 3 shows the participation rates across all students enrolled in grades eligible for an intervention, where participation is measured as receiving at least one session of the treatment. Column 4 shows the percentage of all students targeted for the program who actually participated in it. As for dosage, column 6 shows the approximate average hours that students attended each intervention over the course of the school year, among participating students. For context, we also show here the intended treatment dosage in hours per year, in accordance with each intervention's program design (column 5).

The participation rates for all students in eligible grades for tutoring and small group intervention—shown in panels A and B for math and reading, respectively—ranged from less than 1% to 20%. Relative to tutoring interventions studied in the pre-pandemic literature, some of these programs serve very large numbers of students; whereas just 11% of the programs in Kraft, Schueler, and Falken's (2024) meta-analysis provided tutoring to 400 or more students, half of the tutoring programs in our sample served over 400 students, and most served many more, averaging 2,090 students per program. District A's tutoring program for grades 4 to 8 remarkably delivered tutoring to 10,153 students in reading and 8,461 students in math. For programs with available data on the number or percentage of students targeted for interventions, actual participation rates were typically lower than the share of students program planners intended to serve. With the exception of the math and reading pull-out interventions in District E that served 100% of

TABLE 4

Participation and Dosage of Recovery Interventions

	(1)	(2)	(3)	(4)	(5)	(6)
		Participation			Dosage	
Intervention (Grades)	Sample size	% Targeted in eligible grades	% Treated in eligible grades	% of targeted students treated	Intended dosage in hours per year	Average hours attended per year
A. Tutoring and small group interventions—Math						
District A: Tutoring (4–8)	56,407	28	15	51	30	21.7
District B: Tutoring (5–7)	2,532	—	<1	—	30–60	38.7
District C: Tutoring (K–8)	47,618	22	12	20	9–102	9.4
District C: Tutoring #2 (3–8)	32,274	100	<1	<1	—	11.3–17.7
District E: Pull-out small group (3–8)	8,915	20	20	100	60–80	—
District F: Tutoring (3–8)	15,802	—	4	—	12–36	5.9
District G: Push-in small group (3, 5, 8)	10,565	—	2	—	35	27.1
B. Tutoring and small group interventions—Reading						
District A: Tutoring (4–8)	56,407	32	18	51	30	18.4
District A: Pull-out small group (4–5)	22,713	34	1	4	90	49.9
District B: Tutoring (4–5)	2,266	—	1	—	30–60	31.4
District C: Tutoring (4–8)	26,937	20	10	16	9–102	10.6
District C: Tutoring #2 (3–8)	32,274	100	<1	<1	—	11.3–17.7
District E: Pull-out small group (3–8)	8,915	10	10	100	60–90	—
District F: Tutoring (3–8)	15,802	—	15	—	12–36	8.1
District G: Push-in small group (3, 5, 8)	10,565	—	3	—	35	31.0
District G: Tutoring (3–5)	10,751	—	2	—	18	14.4
C. Other supplemental time interventions—Math						
District A: Extended school year (4–8)	36,987	19	13	61	18	8.0
District B: After-school (4–7)	4,149	—	8	—	14	20.3
District D: Digital learning (4–7)	11,702	29	38	85	30	24.3
District H: After-school (3–8)	1,684	100	24	24	—	4.5–21.8
D. Other supplemental time interventions—Reading						
District A: Extended school year (4–8)	56,407	19	13	61	27	12.0
District B: After-school (4–8)	4,944	—	7	—	14	10.9

(CONTINUED)

TABLE 4 (CONTINUED)

	(1)	(2)	(3)	(4)	(5)	(6)
		Participation			Dosage	
Intervention (Grades)	Sample size	% Targeted in eligible grades	% Treated in eligible grades	% of targeted students treated	Intended dosage in hours per year	Average hours attended per year
District D: Digital learning (4–7)	11,702	28	34	79	30	28.0
District F: Double-dose (6–8)	7,429	27	4	9	124	56.6
District H: After-school (3–8)	1,684	100	24	24	—	4.5–21.8
E. Expert teachers						
Expert teachers (4–8)—Math	36,987	—	26	—	Full school year	Full school year
Expert teachers (4–8)—Reading	56,407	—	19	—	Full school year	Full school year

Note. Sample sizes shown reflect the unrestricted sample of students enrolled in grades eligible for the intervention, meaning students were not required to have MAP Growth scores to be included in this sample. Cells left blank either signify that there are no data available from the district for targeting, intended dosage, or actual dosage for a given intervention. We do not disclose the district that implemented the expert teachers intervention to preserve district anonymity.

targeted students, the remaining tutoring and small group interventions served 51% or fewer of targeted students.

We see similar patterns with regard to dosage. For tutoring and small group interventions, the average hours attended over the course of the year ranged from a low of 5.9 hours (District F Tutoring in math) to a high of 49.9 hours (District A Pull-Out Small Group in reading). Programs varied even more dramatically in their minimum intended dosage per subject, ranging from 9 to 102 hours.[17] Only four programs met their minimum intended dosage (District C Tutoring in math and reading and District B Tutoring in math and reading), but the two District C programs had the lowest minimum intended dosage across programs, 9 hours, and respectively provided just 9.4 and 10.6 hours of tutoring to the average participant.

Although districts generally fell short of their goals both in terms of participation and dosage, there are examples of improvements in these measures across years, 2021–2022 to 2022–2023, for some interventions. Using available data from a subset of these districts and programs from SY 21–22, we find that participation in some of these programs, particularly the larger-scale tutoring programs that scheduled sessions during the school day, generally increased substantially from their first to second year of implementation (see Supplemental Appendix Table A4 in the online version of the journal). Decreases in participation in District F's tutoring program starkly contrast this pattern; our interview notes suggest teachers disliked the virtual program and were less encouraging of students to use it in its second year. Average dosage consistently increased in the second year for these programs.

Panels C and D show participation and dosage statistics for other non-tutoring interventions that provided supplemental instructional time for students in math and reading, respectively. Targeting rates and participation or "take-up" rates are more varied across these programs. For example, District H's After-School programming in math and reading is available to all students, and 24% of students participated. Alternatively, District A's extended school year was targeted at just 19% of students, but over half—61%—of these students participated. Similar to most of the tutoring and small group instruction programs, the programs in this group did not reach all of the targeted students. Interestingly, participation rates exceeded the target rate just for District D's digital learning programs in math and reading, but not all students who participated were targeted. District D's leaders indicated they thought some teachers were using the digital learning program with all of their students, regardless of whether they met the targeting criteria determined by the district. With respect to dosage, the average hours attended ranged widely from 4.5 to 56.6, though for the most part attended hours fell short of intended dosage.

As a classroom-level intervention with mandatory participation (unless a student should switch or opt out of their assigned classroom), the participation and dosage patterns for expert teachers is somewhat distinct. The shares of students assigned to expert teachers across eligible grades are sizable at 26% and 19% for math and reading, respectively. Students do not receive any supplemental instruction time through this intervention; rather, this intervention attempts to accelerate student learning by *replacing* all of a student's instructional time in math or reading over the course of the year with higher-quality instruction.

Intervention Impacts

Tables 5 and 6 report treatment effect estimates from VAMs of tutoring and small group interventions, and other supplemental time interventions, respectively (also displayed in Figures 1 and 2). For each of these tables, column 1 shows the number of students included in the analytic sample used to estimate the intervention's impact; column 2 reports the percent of this analytic sample that received any amount of treatment.[18] In columns 3 to 4, we report the estimated effect of participating in any amount of treatment on math or reading achievement as measured by standard deviations of MAP Growth scores, along with the associated placebo estimate for interventions that are subject-specific.

In addition to estimating the effect of receiving any amount of treatment, we estimate the effect of a single hour of treatment by dividing the estimated coefficients and their standard errors by the average dosage (in hours) received among treated students in the analytic

TABLE 5

Estimated Treatment Effects of Tutoring and Small Group Interventions, Value-Added Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Any participation		Hourly			
Intervention (Grades)	Sample students	% Treated	Point estimate (*SE*)	Placebo estimate (*SE*)	Estimated impact (*SE*)	Placebo estimate (*SE*)	Avg dosage (hours)	Expected effect from research
A. Math interventions								
Overall	101,777	12	0.0323 (0.0387)	—	0.0007 (0.0020)	—	21.0	0.16
District A: Tutoring (4–8)	36,987	18	−0.0061 (0.0129)	−0.0248 (0.0198)	−0.0003 (0.0006)	−0.0011 (0.0009)	22.1	0.16
District B: Tutoring (5–7)	3,627	1	0.2177** (0.0670)	−0.1941*** (0.0258)	0.0059** (0.0018)	−0.0052*** (0.0007)	36.8	0.27
District C: Tutoring (K–8)	39,155	13	−0.0487*** (0.0112)	0.0181 (0.0147)	−0.0051*** (0.0012)	0.0021 (0.0015)	9.6	0.07
District F: Tutoring (3–8)	16,859	3	0.0229 (0.0292)	0.0459 (0.0397)	0.0038 (0.0049)	0.0077 (0.0066)	5.9	0.04
District G: Push-in support (3, 5, 8)	5,149	3	0.0609 (0.0569)	−0.0623 (0.0708)	0.0014 (0.0013)	−0.0014 (0.0017)	43.0	0.32
B. Reading interventions								
Overall	119,935	11	0.0675 (0.0491)	—	0.0015 (0.0016)	—	23.1	0.21
District A: Tutoring (4–8)	31,251	22	−0.0257 (0.0170)	−0.0063 (0.0140)	−0.0013 (0.0009)	−0.0003 (0.0007)	19.1	0.17
District A: Pull-out small group (4–5)	31,251	2	0.3294*** (0.0456)	0.0301 (0.0383)	0.0066*** (0.0009)	0.0006 (0.0008)	49.7	0.44
District B: Tutoring (4–5)	4,184	1	0.2249** (0.0817)	0.1235* (0.0584)	0.0075** (0.0027)	0.0041* (0.0019)	31.4	0.28
District C: Tutoring (4–8)	21,997	11	−0.0196 (0.0158)	0.0148 (0.0131)	−0.0018 (0.0014)	0.0013 (0.0012)	11.0	0.10
District E: Pull-out small group (3–8)	5,480	12	0.0661* (0.0321)	0.0921** (0.0320)	—	—	—	—
District F: Tutoring (3–8)	16,439	13	0.0120 (0.0205)	−0.0148 (0.0246)	0.0015 (0.0026)	−0.0019 (0.0032)	8.4	0.08
District G: Push-in small group (3, 5, 8)	4,915	4	0.0795 (0.0785)	−0.1160 (0.0777)	0.0012 (0.0012)	−0.0017 (0.0012)	66.8	0.59
District G: Tutoring (3–5)	4,915	2	−0.0846 (0.0701)	−0.0777 (0.0659)	−0.0047 (0.0047)	−0.0046 (0.0043)	17.4	0.16

Note. Main effect point estimates show the average effect of receiving any amount of math (or reading) intervention during 2022–2023 on math (or reading) MAP Growth scores in spring 2023. Placebo estimates show the average effect of receiving any amount of these interventions on the opposite subject MAP Growth scores in spring 2023. Covariates in VAMs include participation indicators for other math interventions and reading interventions, prior MAP and state testing (when available) in both math and reading, student demographics, indicators for the calendar week that testing took place for baseline and outcome MAP Growth tests, and school-grade fixed effects. Hourly estimates are calculated by dividing coefficients and standard errors for main and placebo effects by the average dosage, that is, the average number of hours treated students received the intervention over the course of the year. Expected effect from research is calculated by multiplying average dosage by estimated per hour effects of tutoring programs from Nickow et al. (2024) (0.0074 *SD* in math and 0.0089 *SD* in reading). The overall effect of multiple interventions is estimated using a random effects model with REML estimation. The grades shown in the intervention title indicate the grades that a program serves, though the analytic sample for the estimation model may include observations from students in additional grades. Sample students refers to the total number of observations in these analytic samples; % treated refers to the percent of participating students among all students in the analytic sample. All estimates that were significant at $p < .01$ remained significant when accounting for multiple comparisons using the Bonferroni correction. REML = restricted maximum likelihood; VAM = value-added model.

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 6

Estimated Treatment Effects of Other Supplemental Time Interventions, Value-Added Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Any participation		Hourly			
Intervention (Grades)	Sample students	% Treated	Point estimate (*SE*)	Placebo estimate (*SE*)	Estimated impact (*SE*)	Placebo estimate (*SE*)	Avg dosage (hours)	Expected effect from research
A. Math interventions								
Overall	50,441	18	−0.0030 (0.0154)	—	−0.0003 (0.0008)	—	16.8	0.12
District A: Extended school year (4–8)	36,987	14	0.0162 (0.0140)	−0.0083 (0.0183)	0.0019 (0.0017)	−0.0010 (0.0022)	8.4	0.06
District B: After-school (4–7)	3,627	8	−0.0054 (0.0317)	—	−0.0001 (0.0009)	—	20.7	0.15
District D: Digital learning (4–7)	9,827	39	−0.0263 (0.0184)	−0.0142 (0.0182)	−0.0010 (0.0007)	−0.0006 (0.0007)	25.6	0.19
B. Reading interventions								
Overall	61,711	15	0.0073 (0.0132)	—	0.0002 (0.0005)	—	17.2	0.15
District A: Extended school year (4–8)	31,251	13	0.0196 (0.0172)	0.0173 (0.0149)	0.0016 (0.0014)	0.0014 (0.0012)	12.0	—
District B: After-school (4–8)	4,184	7	0.0260 (0.0270)	—	0.0007 (0.0007)	—	11.2	0.10
District D: Digital learning (4–7)	9,837	35	−0.0163 (0.0162)	0.0017 (0.0174)	—	—	28.2	0.25
District F: ELA double dose (1–8)	16,439	1	0.0272 (0.0524)	−0.0588 (0.0373)	0.0004 (0.0008)	−0.0008 (0.0005)	69.5	0.62

Note. Main effect point estimates show the average effect of receiving any amount of math (or reading) intervention during 2022–2023 on math (or reading) MAP Growth scores in spring 2023. Placebo estimates show the average effect of receiving any amount of these interventions on the opposite subject MAP Growth scores in spring 2023. Covariates in VAMs include participation indicators for other math interventions and reading interventions, prior MAP and state testing (when available) in both math and reading, student demographics, indicators for the calendar week that testing took place for baseline and outcome MAP Growth tests, and school-grade fixed effects. Hourly estimates are calculated by dividing coefficients and standard errors for main and placebo effects by the average dosage, that is, the average number of hours treated students received the intervention over the course of the year. Expected effect from research is calculated by multiplying average dosage by estimated per hour effects of tutoring programs from Nickow et al. (2024) (0.0074 *SD* in math and 0.0089 *SD* in reading). The overall effect of multiple interventions is estimated using a random effects model with REML estimation. The grades shown in the intervention title indicate the grades that a program serves, though the analytic sample for the estimation model may include observations from students in additional grades. Sample students refers to the total number of observations in these analytic samples; % treated refers to the percent of participating students among all students in the analytic sample. REML = restricted maximum likelihood; VAM = value-added model.

*$p < .05$. **$p < .01$. ***$p < .001$.

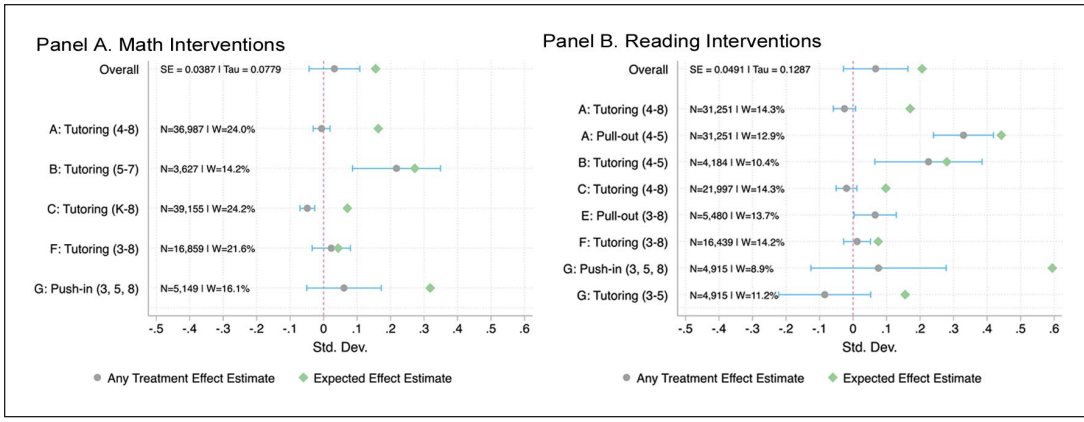

FIGURE 1. *Estimated treatment effects of tutoring and small group instruction interventions.*
Note. We do not display the expected effect estimate for District E's pull-out intervention because data on dosage were not available.

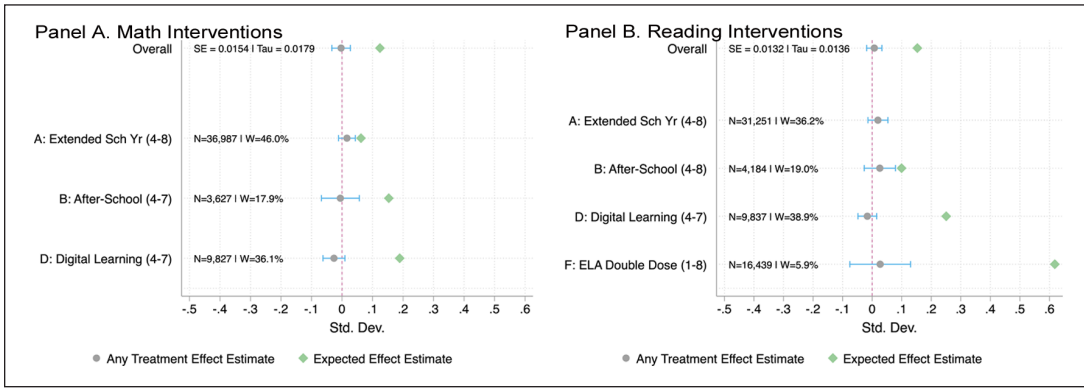

FIGURE 2. *Estimated treatment effects of other supplemental instruction time interventions.*

sample (columns 5–6). Doing so allows us to make comparisons in estimated effectiveness across interventions in a way that does not conflate the dosage received with estimated impact.[19]

Following Carbonari et al. (2024), we put our estimated impacts into context by reporting the effect we would expect to see for each intervention if it were as effective on a per hour basis as high-quality PK–12, pre-pandemic tutoring programs (column 8). To estimate this "expected effect," we use data from Nickow et al.'s (2024) meta-analysis of such programs to derive an estimated per-hour effect of tutoring on math achievement and an estimated per-hour effect on reading achievement. Specifically, we take the average impact of tutoring programs in each subject from that study and divide it by the average number of hours of tutoring offered, where this average is calculated using the same weights as those used in the meta-analysis. Because this rough calculation gives us a benchmark impact per hour of treatment *offered*, rather than received, we adjust these hourly estimates by assuming that students in the meta-analysis studies attended, on average, 93% of the sessions offered, consistent with an overall average national attendance rate of 93% according to NCES data (U.S. Department of Education, 2023). This results in an estimated expected effect *per hour* of tutoring in math of 0.0074 *SD*s and in reading of 0.0089 *SD*s. To calculate the effect we would expect to see given the estimates in Nickow et al. (2024), we multiply the average dosage received by 0.0074 for math

interventions and 0.0089 for reading (these values are reported in column 8 of Tables 5 and 6).

As Table 5 shows, among tutoring and small group interventions, we are unable to detect an overall effect of either math interventions ($\beta = .032$, $p > .05$) or reading interventions ($\beta = .068$, $p > .05$) on student achievement. Among math tutoring and small group interventions, only one program had a positive and significant impact on achievement: District B Tutoring ($\beta = = .218$, $p < .01$). Interestingly, this program stands out both for its relatively high average dosage (37 hours) but also its low treatment percentage—only 1% of students in the analytic sample participated. The placebo effect for District B's tutoring intervention is −0.194 ($p < .001$) and is almost 0.40 *SD* less than the estimated effect of the intervention, suggesting there may have been negative selection bias into the program, and the positive effect estimates may actually be a lower bound. Nevertheless, these results should be interpreted with caution because this program served relatively few students ($n < 30$), and results may be sensitive to small fluctuations in the data.

The only other math tutoring or small group program with a significant estimate is District C Tutoring, though in this case it is significantly negative. This result is somewhat surprising, especially given that the placebo estimate is not statistically different from zero. Relative to the other tutoring and small group instruction programs, District C's program stands out for its greater variation in implementation design (see Supplemental Appendix Table A2 in the online version of the journal): tutoring happened both during school and after school, tutors had a wide variety of qualifications, and the district guidelines around session frequency and duration suggested a student could receive anywhere between 9 and 102 hours of programming. The negative effect may suggest that tutoring was less beneficial for students than participating in their regularly scheduled class period would have been. We speculate this could be the case if the counterfactual for being pulled out of a class to receive tutoring was receiving high-quality small-group instruction with the classroom teacher. It could also be the case that the variation and flexibility in the program design led to inconsistent and less effective programming across schools. It is of course also possible that teachers identified students for interventions using measures not captured by our included prior test scores and covariates. If these students were struggling in math and not in reading, we could observe subject-specific selection bias and a negative point estimate despite a null placebo effect. We cannot say with certainty which (if any) of these explanations are driving the negative findings.

Among reading tutoring and small group interventions, there are two programs with positive and significant impacts that pass, to some degree, the placebo test: District A Pull-Out Small Groups ($\beta = .33$, $p < .001$), and District B Tutoring ($\beta = .22$, $p < .001$). The placebo effect for the first is statistically indistinguishable from zero. For the second, the placebo effect is positive and significant ($\beta = .12$, $p < .05$) though its magnitude is a full 0.1 standard deviation below that of the main effect estimate, suggesting that while some positive selection may be resulting in an overestimate of the effect, it may not be sufficient to account for the full impact. Similar to the math tutoring and small group intervention for which we detected a positive effect, both District A Pull-Out Small Groups and District B Tutoring had notably low participation rates among their analytic samples (2% and 1%, respectively) and relatively high average dosages (49.7 and 31.4 hours, respectively).

Table 6 shows that, as with tutoring and small group interventions, we are unable to detect overall effects of districts' other supplemental time interventions on either math achievement ($\beta = -.003$, $p > .05$) or reading achievement ($\beta = .007$, $p > .05$). Additionally, we are unable to detect effects of any of the three math interventions or four reading interventions individually. The magnitude of the estimates for this set of interventions are all generally small (the largest estimate being for District F ELA Double-Dose, $\beta = .027$), suggesting that it is not simply a case of imprecision that prevents us from detecting impacts of these interventions.

In Table 7, we report estimated treatment effects and placebo effects of expert teachers on math and reading achievement. The expert teacher intervention is fundamentally different from the other interventions studied because students in the control group are necessarily assigned to non-expert teachers and are thus also

TABLE 7

Estimated Treatment Effects of Expert Teachers, Value-Added Models

	(1)	(2)	(3)	(4)
			Any participation	
Intervention (Grades)	Sample students	% Treated	Point estimate (*SE*)	Placebo estimate (*SE*)
Expert teachers in Math (4–8)	36,987	32	0.0571*** (0.0135)	0.0005 (0.0138)
Expert teachers in ELA (4–8)	31,251	20	0.1083*** (0.0140)	0.0230 (0.0116)

Note. We do not disclose the district that implemented the expert teachers intervention to preserve district anonymity. Main effect point estimates show the average effect of receiving any amount of math (or reading) intervention during 2022–2023 on math (or reading) MAP Growth scores in spring 2023. Placebo estimates show the average effect of receiving any amount of these interventions on the opposite subject MAP Growth scores in spring 2023. Covariates in VAMs include participation indicators for other math interventions and reading interventions, prior MAP and state testing (when available) in both math and reading, student demographics, indicators for the calendar week that testing took place for baseline and outcome MAP Growth tests, and school-grade fixed effects. The grades shown in the intervention title indicate the grades that a program serves, though the analytic sample for the estimation model may include observations from students in additional grades. Sample students refers to the total number of observations in these analytic samples; % treated refers to the percent of participating students among all students in the analytic sample. All estimates that were significant at $p < .01$ remained significant when accounting for multiple comparisons using the Bonferroni correction. VAM = value-added model.

*$p < .05$. **$p < .01$. ***$p < .001$.

affected by the treatment. The treatment-control contrast, therefore, is assignment to an expert versus a non-expert teacher in a given grade and subject, rather than assignment to an expert teacher versus "business-as-usual" (i.e., having a chance of being assigned to an expert teacher or a non-expert teacher). The intervention also does not provide students with any additional instruction time beyond what they would receive with a non-expert teacher. For these reasons, we do not provide hourly estimates of the program's impacts or a comparison to benchmark estimates of the impact of tutoring (i.e., the expected effect from research). Notably, we find positive and significant impacts of having an expert teacher (as opposed to a non-expert teacher) in math ($\beta = .057$, $p < .001$) and in reading ($\beta = .108$, $p < .001$). The placebo tests for both subjects support the claim that these detected impacts are not the result of selection into the expert teacher classrooms. The placebo estimate for math is non-significant and very close to zero, and for reading is non-significant with a magnitude that is far lower than that of the main point estimate ($\beta_{placebo} = .023$ vs. $\beta_{main} = .108$). As opposed to the handful of tutoring programs with positive impacts, the expert teachers intervention stands out as the one program not limited to a considerably small group of students, with participation rates among the analytic samples of 32% in math and 20% in reading.

Finally, we describe the results of the fuzzy RD analysis that we conducted for District A Tutoring in math and reading, and District D Digital Learning in math and reading. As discussed below, although we have concerns about the validity of the RD analysis, we nevertheless report the first- and second-stage fuzzy RD estimates across bandwidths of 0.5 and 1 *SD* of the running variable, as shown in Table 8. We find the first stage is statistically strong ($p < .001$) for each intervention. For District A in particular, the likelihood of receiving tutoring jumps by ~77% at the eligibility threshold for math and by ~83% for reading. In District D, the eligibility threshold is less predictive of treatment receipt, though still statistically significant, with a ~35% jump for math and ~46% for reading. However, in all cases, the second stage results show no statistically significant discontinuity in test scores at the eligibility threshold. These null findings are consistent with those from our VAMs of the same interventions, but the RD estimates are less precise (i.e., have larger standard errors).

TABLE 8

Estimated Treatment Effects From Regression Discontinuity Models

Intervention subject	Bandwidth (*SD*)	Model	*N*	Point estimate	*SE*
District A: Tutoring					
Math	0.5	First stage	9,154	0.7715***	0.0114
	0.5	Second stage	9,154	0.0081	0.0433
	0.1	First stage	16,410	0.7743***	0.0088
	0.1	Second stage	16,410	−0.0314	0.0270
Reading	0.5	First stage	7,212	0.8300***	0.0124
	0.5	Second stage	7,212	−0.0098	0.0533
	1	First stage	14,028	0.8271***	0.0095
	1	Second stage	14,028	0.0166	0.0381
District D: Digital learning					
Math	0.5	First stage	2,989	0.3329***	0.0384
	0.5	Second stage	2,989	0.1272	0.1342
	1	First stage	5,439	0.3662***	0.0088
	1	Second stage	5,439	0.0233	0.0877
Reading	0.5	First stage	2,479	0.4647***	0.0511
	0.5	Second stage	2,479	0.0653	0.1453
	1	First stage	5,049	0.4595***	0.0397
	1	Second stage	5,049	0.0236	0.0987

Note. The outcome of the First Stage estimates is the probability of being treated, while the outcome of the Second Stage is the norms-standardized spring MAP test. For both districts, the running variable is the standardized test score that determined eligibility for the intervention, centered at the eligibility threshold.
*$p < .05$. **$p < .01$. ***$p < .001$.

We interpret the RD results with caution, as the validity checks presented in Supplemental Appendix C (available in the online version of this article) reveal statistically significant discontinuities in the density of the running variable around the cutoff for both subjects in District A and for math in District D. While this pattern is suggestive of sorting or manipulation near the cutoff, there is no clear mechanism for schools or students to influence test scores in this way. Additional sensitivity analyses, including alternative bandwidths, weighting schemes, and model specifications, yield consistent first-stage estimates and similarly null second-stage results (see Supplemental Appendix C in the online version of the journal). However, due to the imprecision in the estimates and concerns raised by the failed density tests, we consider the VAM results more reliable than the RD results for these interventions.

For the handful of interventions where we do find significant evidence of positive impacts using VAMs, existing research provides a guide for interpreting effect sizes. The three tutoring or small group interventions with positive impacts (District B Tutoring in math, District A Pull-Out Small Group in reading, and District B Tutoring in reading) all had estimated effects that were smaller in magnitude than what would be expected based on per hour estimates from Nickow et al., 2024; they fell short of these estimates by approximately 19% to 25%. Kraft (2020) reviews 750 RCTs that estimate the effect of educational interventions on achievement and proposes empirical benchmarks that classify effect sizes below 0.05 *SD* as small, between 0.05 and 0.20 *SD* as medium, and equal to or greater than 0.20 *SD* as large. Taken together, these benchmarks suggest that those three tutoring or small group interventions had large impacts on those students who participated. The effect sizes of the expert teachers program, in comparison, were on the lower side of medium.

Conclusion

In this study, we estimate the effects of academic COVID recovery interventions on student achievement during the 2022–2023 school year in eight large districts. We find few programs significantly impacted student achievement. Just two tutoring programs effectively improved students' reading achievement (+0.22 to +0.33 *SD*), and only one program improved students' math achievement (+0.22 *SD*). We estimate a significant negative impact of District C's math tutoring program (−0.05 *SD*), suggesting the program had less benefit for student achievement than the typical instruction students were receiving during that time.

The three tutoring programs with positive impacts were intensive, averaging over 30 hours of instruction per student. The hourly effects of these intensive programs were similar to, or just below, those found in previous RCTs (Nickow et al., 2024), resulting in overall effects that are large for educational interventions (Kraft, 2020). However, intensity alone did not guarantee success. Two push-in small group instruction programs and an ELA double-dose program (all of which served less than 4% of students in eligible grades) showed no detectable effects despite providing students with 43, 67, and 70 hours of supplemental instruction over the year, respectively.

Though they were effective, the three tutoring interventions' positive impact was constrained by their limited scope. Each was so small—serving just 1% to 2% of eligible students—that they were unlikely to make significant contributions to their district's overall academic recovery or serve as a model for large-scale interventions. More concerning, the large-scale interventions in our study—which reached between 7% and 39% of students in grades 4 to 8—failed to produce significant improvements in student achievement. These results broadly align with Kraft, Edwards, and Cannata's (2024) meta-analysis showing that pre-pandemic tutoring program effects attenuate substantially with program size, though the large interventions in the present study are still less effective than those in the meta-analysis that serve over 1,000 students.

We do find a notable departure from the null effects described above: the expert teacher intervention. Expert teachers served 32% of eligible students in math and 20% in ELA. We find that being assigned to an expert teacher significantly improved student achievement by +0.06 *SD* in math and +0.11 *SD* in reading, relative to being assigned to a non-expert teacher. While this intervention may have benefited the students assigned to expert teachers, it would not be expected to raise achievement overall if it was offsetting relative losses for students assigned to non-expert teachers. It is also difficult to assess the "dosage" of this program, as the program intends to functionally replace students' instructional time with higher quality time, as opposed to supplementing it. For these reasons, we cannot directly compare the effects of expert teachers and the other interventions in this study. Nevertheless, these results underscore the potential to accelerate student learning by focusing on teacher quality and maximizing existing class time.

With the influx of ESSER funds, districts had more capacity than usual to expand staffing or hire contractors to deliver interventions. Our findings suggest, however, they often struggled to deliver intensive, effective programs to all the students who needed support recovering from pandemic-related declines in achievement. How do we reconcile low participation rates in effective programs and otherwise null results of interventions with the moderate improvement in district-level achievement from spring 2022 to spring 2023 state test scores reported by Fahle et al. (2024)?

There are at least two possible explanations. One is that our analysis does not include district-wide recovery efforts (e.g., Tier 1 interventions such as instructional coaches, new curriculum). To the extent that districts hired more teachers, paraprofessionals and school counselors that equally benefited all students, it would not show up in our analysis. And because we compare treated students to untreated students to estimate intervention impacts, we also cannot capture any district-wide effects of targeted interventions (e.g., if interventions had positive spillover effects).

A second possibility is that the state test scores used in Fahle et al. (2024) reflect score inflation and not increases in real learning. Since there was no NAEP test in 2023, their estimates assumed that the NAEP equivalents estimated for state proficiency thresholds in 2022 applied in

2023. When a new NAEP is released in 2024, and the researchers recalibrate state proficiency thresholds, it is possible some of the increase will be revised downward. However, the fact that Dewey et al. (2024) and Goldhaber and Falken (2024) both find that districts that received larger ESSER grants per student also saw larger improvements on the NAEP test, makes the first explanation more likely: that the improvement that we saw between 2022 and 2023 was due to district-wide efforts affecting all students, rather than targeted catchup efforts.[20]

If districts want to address achievement gaps and help the students most harmed by school closures, they will need to improve their targeted catchup efforts. Our study suggests these efforts, especially at scale, may present a core dilemma: a trade-off between participation rates and program intensity. Pre-pandemic evaluations of small-scale interventions, despite their importance, provide little insight into large-scale implementation. Like a chef adapting a sophisticated recipe they prepared for a small dinner party to serve a large banquet, decision-makers need guidance on how they can modify effective interventions to target larger groups of students without substantially diminishing their effectiveness. The use of technology to support and standardize these programs is a potentially promising path forward: As noted earlier, three recent RCTs of pandemic-era tutoring interventions that used virtual tutors and/or digital learning tools all had significant positive effects on achievement and reached sizeable groups of students (respectively 420, 2,060, and 959 students; Bhatt et al., 2024; Cortes et al., 2025; Ready et al., 2024).

Future research needs to do much more to develop and test ideas for effectively scaling up interventions in ways that balance cost, participation, and impact. Specifically, school systems and policymakers need better evidence on which intervention features (and combinations of features) accelerate student learning, for which students, in what contexts, and at what cost (Kohlmoos & Steinberg, 2024). As pandemic-impacted students continue to progress through K–12 education with limited evidence of academic recovery (Curriculum Associates, 2023, 2024; Fahle et al. 2024; Lewis & Kuhfeld, 2023, 2024), the need for action is urgent.

Acknowledgment

We thank Jazmin Isaacs for her efforts in cleaning the NWEA data.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by funding from Accelerate (256843-5124839) and AIR Equity Initiative: Educational Equity through Policy Implementation (0640002302).

ORCID iDs

Elise Dizon-Ross https://orcid.org/0000-0002-0396-1523
Dan Goldhaber https://orcid.org/0000-0003-4260-4040
Andrew McEachin https://orcid.org/0000-0002-5113-6616
Emily Morton https://orcid.org/0000-0003-3662-8556
Atsuko Muroga https://orcid.org/0000-0003-4431-8590

Notes

1. Fahle et al. (2024) estimate the amount of academic recovery needed to return to pre-pandemic achievement levels in math shrank by about a third from spring 2022 to spring 2023, and the amount of recovery needed in ELA shrunk by about by about a quarter.

2. We exclude from our analysis several interventions that districts identified as academic recovery interventions that served a subset of students but were used as part of core instruction, such that the counterfactual to receiving treatment (and what any estimated effect would represent) was unknown.

3. Some interventions served students in grades beyond K–8, but we limited the scope of our study to interventions serving students in K–8 because NWEA MAP testing beyond 8th grade was uncommon in our districts.

4. For example, Jacob and Lefgren (2009) find that retaining low-performing eighth grade students increases the likelihood that these students later drop-out of high school.

5. More recently, Kraft and Lovison (2024) provide experimental evidence that finds 1:1 online tutoring is more effective than 3:1 online tutoring.

6. The estimate of the average instructional dosage for this program was retrieved through correspondence with the study authors and tutoring provider.

7. For more about the R2R project, including other research findings, see https://caldercenter.org/road-covid-recovery.

8. $z(Y_{igst}) = (Y_{igt} - \bar{Y}_{gt})/SD(Y_{gt})$

9. We co-created notes during these conversations to maximize transparency and the accuracy of the information we collected. A notetaker shared their screen with participants and shared their notes with participants following the interview. We encouraged participants to correct any information that did not represent their understanding of the intervention's implementation.

10. The grade levels eligible for the expert teacher intervention, 4 to 8, had departmentalized instruction, such that students in these grades could be assigned to an expert teacher in math and/or reading.

11. District C's extended year intervention was also excluded from the analysis because the intervention was delivered at the school-level to a limited number of schools, limiting the statistical power to detect effects.

12. See Thum and Kuhfeld (2020) for details on NWEA's pre-pandemic norms.

13. In some districts, intervention eligibility is determined fully or in part by these earlier MAP scores and/or by other test scores such as those from state standardized tests. In these cases, we also include a cubic polynomial function of the relevant test scores.

14. We additionally estimate versions of these models where we cluster standard errors at the school level, and where we calculate robust (un-clustered) standard errors. We arrive at consistent findings with respect to the statistical significance of our results. The one exception is District E Pull-Out Small Group in reading, the estimate for which becomes marginally significant ($p = .07$) when clustering standard errors at the school level.

15. See also Supplemental Appendix Tables A5 and A6 (available in the online version of this article), which describe the student characteristics and prior test scores of treated and untreated students for each intervention.

16. Other interventions listing eligibility as a "MAP score range" appear to have used additional criteria, as assignment did not follow a strict cutoff.

17. For comparison, the average dosage of tutoring offered among the studies included in Nickow et al.'s (2024) meta-analysis were 39 hours per year in math and 35 hours per year in reading. The tutoring programs in Kraft, Schueler, and Falken's (2024) meta-analysis offered an average of 35 hours of tutoring per year per subject. These estimates do not take into account the percent of sessions actually attended by students.

18. Because the participation rates shown in Tables 5 and 6 use the students included in the value-added analytic sample as the denominator, rates may differ from those shown in Table 4, which reports participation rates among all students in eligible grades in the district.

19. This approach follows Carbonari et al. (2024). We estimate hourly effects in this way, rather than including a continuous measure of hours of treatment received in the VAMs, out of concern that at an individual student level, the amount of intervention received is likely to be endogenous. For instance, we might guess that a student struggling more may be more likely to participate in a large number of sessions; or alternatively, a highly motivated student may be more likely to participate in a large number of sessions. To avoid potential bias from the potential for this type of dosage endogeneity, we instead divide by the average number of hours received across all students. Note that the intent of this approach is to provide comparable estimates, rather than to provide meaningful estimates on the internal margin of treatment, which we do not model in this paper.

20. Unfortunately, there is very little detailed information regarding how districts spent their ESSER money, such that it is hard to say which types of investments were driving recovery.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering?. *The Quarterly Journal of Economics*, *138*(1), 1–35.

Arthur, A. M., & Davis, D. L. (2016). A pilot study of the impact of double-dose robust vocabulary instruction on children's vocabulary growth. *Journal of Research on Educational Effectiveness*, *9*(2), 173–200.

Barry, S. S., & Sass, T. R. (2022). *The impact of a 2021 Summer School Program on student achievement*. Georgia Policy Labs. https://doi.org/10.57709/FAJ9-8597

Bettinger, E., Fairlie, R., Kapuza, A., Kardanova, E., Loyalka, P., & Zakharov, A. (2023). Diminishing marginal returns to computer-assisted learning. *Journal of Policy Analysis and Management*, *42*(2), 552–570.

Bhatt, M. P., Guryan, J., Khan, S. A., LaForest-Tucker, M., & Mishra, B. (2024). *Can technology facilitate scale? Evidence from a randomized evaluation of high dosage tutoring* (NBER Working Paper No. w32510). National Bureau of Economic Research.

Callen, I., Carbonari, M. V., DeArmond, M., Dewey, D., Dizon-Ross, E., Goldhaber, D., Isaacs, J., Kane, T. J., Kuhfeld, M., McDonald, A., McEachin, A., Morton, E., Muroga, A., & Staiger, D. O. (2023). *Summer school as a learning loss recovery strategy after COVID-19: Evidence from Summer 2022* (Working Paper No. 291-0823). National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Carbonari, M. V., Davison, M., DeArmond, M., Dewey, D., Dizon-Ross, E., Goldhaber, D., Hashim, A., Kane, T. J., McEachin, A., Morton, E., Muroga, A., Patterson, T., & Staiger, D. O. (2024). *The challenges of implementing academic COVID recovery interventions: Evidence from the Road to Recovery Project* (Working Paper No. 275-0624-2). National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Checkoway, A., Gamse, B., Velez, M., & Linkow, T. (2013). *Evaluation of the Massachusetts expanded learning time (ELT) initiative: Final study findings*. Society for Research on Educational Effectiveness.

Cortes, K., Kortecamp, K., Loeb, S., & Robinson, C. (2025). A scalable approach to high-impact tutoring for young readers. *Learning and Instruction*, *95*, 102021.

Curriculum Associates. (2023). *State of student learning in 2023*. https://cdn.bfldr.com/LS6J0F7/at/x8v8wp2c6j4s4wttsw2nwphb/ca-state-of-student-learningtechnical-report-2023.pdf

Curriculum Associates. (2024). *State of student learning in 2024*. https://cdn.bfldr.com/LS6J0F7/at/wxj37b648k5bkwvf3vrrgj8/2024-State-of-Student-Learning-Research-Awareness-Technical-Report.pdf

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Dewey, D., Fahle, E., Kane, J., Reardon, S., & Staiger, D. (2024). *Federal pandemic relief and academic recovery*. Center for Education Policy Research at Harvard University.

Diliberti, M. K., & Schwartz, H. L. (2022). *Districts continue to struggle with staffing, political polarization, and unfinished instruction*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA956-13.html

Doty, E., Kane, T. J., Patterson, T., & Staiger, D. O. (2022). What do changes in state test scores imply for later life outcomes? (NBER Working Paper No. 30701). National Bureau of Economic Research.

Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). *Education technology: An evidence-based review* (NBER Working Paper No. 23744). National Bureau of Economic Research.

Fahle, E. M., Kane, T. J., Patterson, T., Reardon, S. F., Staiger, D. O., & Stuart, E. A. (2023). *School district and community factors associated with learning loss during the COVID-19 pandemic*. Center for Education Policy Research at Harvard University.

Fahle, E. M., Kane, T. J., Reardon, S. F., & Staiger, D. O. (2024). *The first year of pandemic recovery: A district-level analysis*. Center for Education Policy Research at Harvard University.

Goldhaber, D., & Falken, G. (2024). *ESSER and student achievement: Assessing the impacts of the largest one-time federal investment in K12 schools* (CALDER Working Paper No. 301-0624). National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., & Staiger, D. O. (2023). The educational consequences of remote and hybrid instruction during the pandemic. *American Economic Review: Insights*, *5*(3), 377–392.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201–209.

Hanushek, E. A., & Strauss, B. (2024). *A global perspective on US learning losses*. Hoover Institution. https://www.hoover.org/sites/default/files/research/docs/HanushekStrauss_WebreadyPDF_240229.pdf

Hedges, L. V. (1983). Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology*, *36*, 123–131. https://doi.org/10.1111/j.2044-8317.1983.tb00768.x

Hill, H. C., & Erickson, A. (2019). Using implementation fidelity to aid in interpreting program impacts: A brief review. *Educational Researcher*, *48*(9), 590–598.

Isaacs, J., Kuhfeld, M., & Lewis, K. (2023). *Technical appendix for: Education's long COVID: 2022–23 Achievement data reveal stalled progress towards pandemic recovery*. NWEA. https://www.nwea.org/uploads/Tech-appendix-July-2023-Final.pdf

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33–58.

Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from Kindergarten to Grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, *83*(3), 386–431.

Kohlmoos, L., & Steinberg, M. P. (2024). *Contextualizing the impact of tutoring on student learning: Efficiency, cost effectiveness, and the known unknowns* [Research report] Accelerate. https://accelerate.us/efficiency-and-cost-effectiveness

Kraft, M. A. (2015). How to make additional time matter: Integrating individualized tutorials into an extended day. *Education Finance and Policy*, *10*(1), 81–116.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

Kraft, M. A., Edwards, D. S., & Cannata, M. (2024). *The scaling dynamics and causal effects of a district-operated tutoring program* (EdWorkingPaper No. 24-1030). Annenberg Institute at Brown University. https://doi.org/10.26300/zcw7-4547

Kraft, M. A., Schueler, B. E., & Falken, G. (2024). *What impacts should we expect from tutoring at scale? Exploring meta-analytic generalizability* (EdWorkingPaper No. 24-1031). Annenberg Institute at Brown University. https://edworkingpapers.com/ai24-1031

Kraft, M. A., & Lovison, V. S. (2024). *The effect of student-tutor ratios: Experimental evidence from a Pilot Online Math Tutoring Program* (EdWorkingPaper No. 24-976.) Annenberg Institute at Brown University.

Kraft, M. A., & Novicoff, S. (2024). Time in school: A conceptual framework, synthesis of the causal research, and empirical exploration. *American Educational Research Journal*, *61*(4), 724–766.

Kuhfeld, M., Diliberti, M., McEachin, A., Schweig, J., & Mariano, L. T. (2023). *Typical learning for whom? Guidelines for selecting benchmarks to calculate months of learning* (Research brief). NWEA.

Lewis, K., & Kuhfeld, M. (2022). *Progress toward pandemic recovery: Continued signs of rebounding achievement at the start of the 2022–23 school year* (Research brief). NWEA.

Lewis, K., & Kuhfeld, M. (2023). *Education's long COVID: 2022–23 Achievement data reveal stalled progress toward pandemic recovery* (Research brief). NWEA.

Lewis, K., & Kuhfeld, M. (2024). *Recovery still elusive: 2023–24 Student achievement highlights persistent achievement gaps and a long road ahead* (Research brief). NWEA.

Lynch, K., An, L., & Mancenido, Z. (2023). The impact of summer programs on student mathematics achievement: A meta-analysis. *Review of Educational Research*, *93*(2), 275–315.

Makori, A., Burch, P., & Loeb, S. (2024). *Scaling high-impact tutoring: School level perspectives on implementation challenges and strategies* (EdWorkingPaper No. 24-932). Annenberg Institute at Brown University.

McCombs, J. S., Pane, J. F., Augustine, C. H., Schwartz, H. L., Martorell, P., & Zakaras, L. (2014). *Ready for fall? Near-term effects of voluntary summer learning programs on low-income students' learning opportunities and outcomes*. RAND Corporation. https://doi.org/10.7249/RR815

McCombs, J. S., Whitaker, A. A., & Yoo, P. (2017). *The value of out-of-school time programs*. RAND Corporation. https://www.rand.org/pubs/perspectives/PE267.html

McEachin, A., & Atteberry, A. (2017). The impact of summer learning loss on measures of school performance. *Education Finance and Policy*, *12*(4), 468–491.

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, *61*(1), 74–107.

Nomi, T. (2015). "Double-dose" English as a strategy for improving adolescent literacy: Total effect and mediated effect through classroom peer ability change. *Social Science Research*, *52*, 716–739.

Nomi, T., & Allensworth, E. M. (2013). Sorting and supporting: Why double-dose algebra led to better test scores but more course failures. *American Educational Research Journal*, *50*(4), 756–788.

Opper, I., & Özek, U. (2024). *A global regression discontinuity design: Theory and application to grade retention policies* (CESifo Working Paper No. 10972.) Center for Economic Studies.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics*, *37*(2), 187–204.

Özek, U. (2021). The effects of middle school remediation on postsecondary success: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *203*, 104518.

Özek, U., & Mariano, L. T. (2023). *Think again: Is grade retention bad for kids?*. Thomas B. Fordham Institute.

Pollard, C., Lu, A., Zandieh, A., Robinson, C. D., Loeb, S., & Waymack, N. (2024). *Implementation of the OSSE high impact tutoring initiative: First year report school year 2022–2023* (Research report). National Student Support Accelerator. https://studentsupportaccelerator.org/briefs/implementation-osse-high-impact-tutoring-initiative

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). Russell Sage Foundation.

Ready, D. D., McCormick, S. G., & Shmoys, R. J. (2024). *The effects of in-school virtual tutoring on student reading development: Evidence from a short-cycle randomized controlled trial* (EdWorkingPaper No. 24-942). Annenberg Institute at Brown University.

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & Saliba, J. (2024). *Stanford Education Data Archive* (Version 5.0). The Educational Opportunity Project at Stanford University. https://purl.stanford.edu/cs829jn7849

Robinson, C. D., Bisht, B., & Loeb, S. (2022). *The inequity of opt-in educational resources and an intervention to increase equitable access* (EdWorkingPaper No. 22-654). Annenberg Institute at Brown University.

Sacerdote, B. (2012). When the saints go marching out: Long-term outcomes for student evacuees from Hurricanes Katrina and Rita. *American Economic Journal: Applied Economics*, *4*(1), 109–135.

The White House. (2024, January 17). *FACT SHEET: Biden-Harris administration announces improving student achievement agenda in 2024*. https://www.whitehouse.gov/briefing-room/statements-releases/2024/01/17/fact-sheet-biden-harris-administration-announces-improving-student-achievement-agenda-in-2024/

Thum, Y. M., & Kuhfeld, M. (2020, April). *NWEA 2020 MAP growth: Achievement status and growth norms—Tables for students and schools*. NWEA. https://teach.mapnwea.org/impl/NormsTables.pdf

U.S. Department of Education. (2021). *Education in a pandemic: The disparate impacts of COVID-19 on America's students*. U.S. Department of Education, Office of Civil Rights.

U.S. Department of Education. (2022). *National Assessment of Educational Progress (NAEP) 2022 long-term trend assessment results: Reading and mathematics*. Institute of Education Sciences, National Center for Education Statistics. https://www.nationsreportcard.gov/highlights/ltt/2022/

U.S. Department of Education. (2023). *Digest of education statistics: 2023*. Institute of Education Sciences, National Center for Education Statistics.

Authors

MARIA V. CARBONARI, MA, is a doctoral student at the Graduate School of Education at the University of Pennsylvania. Her research focuses on topics related to education policy and the economics of education.

MICHAEL DEARMOND, PhD, is director of policy at CALDER at the American Institutes for Research. His research focuses on educational governance, bureaucratic reform, and policy implementation.

DANIEL DEWEY, MA, is a senior research analyst at the Center for Education Policy Research at Harvard University. His research focuses on pandemic learning loss, interventions, recovery, and long-term outcomes of charter schooling.

ELISE DIZON-ROSS, PhD, is a researcher at CALDER at the American Institutes for Research. Her research examines the impacts of economic inequality and access to basic needs on student outcomes and the education sector, from K–12 through higher education.

DAN GOLDHABER, PhD, is the director of CALDER at the American Institutes for Research and the director of the Center for Education Data & Research at the University of Washington. His research focuses on issues of educational productivity and reform at the K–12 level; the broad array of human capital policies that influence the composition, distribution, and quality of teachers in the workforce; and connections between students' K–12 experiences and postsecondary outcomes.

THOMAS J. KANE, PhD, is the Walter H. Gale Professor of Education and Faculty Director of the Center for Education Policy Research at the Harvard Graduate School of Education. His research focuses on K–12 and higher education, covering topics such as the design of school accountability systems, teacher recruitment and retention, financial aid for college, race-conscious college admissions, and the earnings impacts of community colleges.

ANNA MCDONALD, BA, is a research associate at CALDER at the American Institutes for Research. Her research focuses on evaluating the impacts of educational interventions, including summer learning and tutoring.

ANDREW MCEACHIN, PhD, is a senior research director at the ETS Research Institute. His work focuses on helping policymakers and educators make informed decisions about the design and implementation of educational policies.

EMILY MORTON, PhD, is a lead research scientist at NWEA. Her research focuses on estimating the effects of K–12 education policies and programs related to instructional time and learning environments on student outcomes.

ATSUKO MUROGA, PhD, is a postdoctoral scholar at the Center for Education Policy Research at Harvard University. Her research focuses on early childhood education, child development, early reading education, wraparound services, and university-school-community partnerships.

ALEJANDRA SALAZAR, BA, is a research associate at CALDER at the American Institutes for Research.

Her research focuses on the economics of education and the application of computational methods to educational data.

DOUGLAS O. STAIGER, PhD, is the John Sloan Dickey Third Century Professor in the Department of Economics at Dartmouth College. His research interests include the economics of education, the economics of healthcare, and statistical methods.

Manuscript received August 23, 2024
First revision received August 29, 2025
Second revision October 15, 2025
Accepted October 28, 2025