Impacts of Academic Recovery Interventions on Student Achievement in 2022-23

Maria V. Carbonari
Michael DeArmond
Daniel Dewey
Elise Dizon-Ross
Dan Goldhaber
Thomas J. Kane
Anna McDonald
Andrew McEachin
Emily Morton
Atsuko Muroga
Alejandra Salazar
Douglas O. Staiger

July 2024

WORKING PAPER No. 303-0724

CALDER
National Center for Analysis of
Longitudinal Data in Education Research

AIR®

Impacts of Academic Recovery Interventions on Student Achievement in 2022-23

Maria V. Carbonari
Harvard University

Daniel Dewey
Harvard University

Thomas J. Kane
Harvard University

Atsuko Muroga
Harvard University

Michael DeArmond
American Institutes for Research / CALDER

Elise Dizon-Ross
American Institutes for Research / CALDER

Dan Goldhaber
American Institutes for Research / CALDER
University of Washington

Anna McDonald
American Institutes for Research / CALDER

Emily Morton
American Institutes for Research / CALDER

Alejandra Salazar
American Institutes for Research / CALDER

Andrew McEachin
NWEA

Douglas O. Staiger
Dartmouth College

Contents

Acknowledgments

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders or the institutions to which the authors are affiliated. Any errors are attributable to the authors.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

Impacts of Academic Recovery Interventions on Student Achievement in 2022-23

Maria V. Carbonari, Michael DeArmond, Daniel Dewey, Elise Dizon-Ross, Dan Goldhaber, Thomas J. Kane, Anna McDonald, Andrew McEachin, Emily Morton, Atsuko Muroga, Alejandra Salazar, & Douglas O. Staiger

CALDER Working Paper No. 303-0724
July 2024

Abstract

The COVID-19 pandemic devastated student achievement, with declines rivaling those after Hurricane Katrina. These losses widened achievement gaps between historically marginalized students and their peers. Three years later, achievement remains behind pre-pandemic levels for many students. This paper examines 2022-23 academic recovery efforts across eight districts, including tutoring, small group instruction, after-school, extended year, double-dose, digital learning, and expert teacher interventions. Across 22 math and reading interventions, most were delivered to fewer students and for less time than planned. We find positive effects for one tutoring program on math scores and two tutoring programs on reading scores, ranging from 0.22 to 0.33 SD. Each of these programs served a very small share of the district's students and was unlikely to play a major role in district-wide academic recovery. Finally, we find that having an "expert" teacher with high evaluation scores as opposed to a non-expert teacher significantly improves student achievement by 0.06 SD in math and 0.11 SD in reading. While highlighting the promise of intensive academic interventions, our findings underscore the challenges districts face in scaling such interventions to match their recovery needs. The field needs better evidence regarding successful implementation of large-scale interventions.

1. Introduction

The COVID-19 pandemic had a significant negative impact on student achievement, with nationwide average declines comparable to those observed after Hurricane Katrina (approximately 0.17 standard deviations (SD); Sacerdote, 2012). Pandemic-related disruptions to public schooling and other social services especially affected students from historically marginalized groups. Achievement gaps widened, arguably undoing nearly two decades of progress toward educational equity in the United States. (United States Department of Education, 2021, 2022). As of the spring of 2023, three 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., 2024; Lewis & Kuhfeld, 2023).

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, included 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). 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 early 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, from scheduling problems, staffing shortages and absenteeism, to inadequate central office capacity (Carbonari et al., 2024; Markori 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

2021-22 school year (Barry & Sass, 2022; Carbonari et al., 2024; Callen et al., 2023; Robinson et al., 2022).

By the summer of 2022, however, evidence of improvement was starting to emerge. Analyzing the academic progress of students who attended summer school in 2022 in eight school districts, Callen et al. (2023) found a positive impact for summer school on math test achievement (+0.03 SD), but not in reading. Using data from state tests, Fahle et al. (2024) find 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 (Dewey et al., 2024; Goldhaber & Falken, 2024). Despite this progress, recent studies using interim assessment data from NWEA (Lewis & Kuhfeld, 2023) and Curriculum Associates (Curriculum Associates, 2023) show that spring 2023 test scores remained far below pre-pandemic levels and that historically marginalized students are still the furthest behind.

The stakes surrounding students' academic recovery remain high. Hanushek (2023), for example, estimates that unremedied declines in students' test scores could translate to average lifetime earnings reductions of 2 to 9 percent for students and a 3.5 percent decrease in economic growth, totalling $31 trillion. Doty et al. (2023) forecast smaller (but still large) impacts when limited to individual students' lifetime earnings ($900 billion). Importantly, the likely downstream impacts are even more severe for students of color and economically disadvantaged students, who suffered disproportionately from the pandemic's disruptions. Four years after

[1] Fahle et al. (2024) estimate the amount of academic recovery needed to return to pre-pandemic achievement levels in math shrunk 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.

COVID upended schools nationwide, learning acceleration and academic recovery remain critical issues for the United States' economy and social equality (The White House, 2024).

This paper extends our prior analyses of COVID recovery (Carbonari et al., 2024; Callen et al., 2023) 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 three organizations, the American Institutes for Research, Harvard University, and NWEA, a nonprofit testing company.

We primarily 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"). We also examine a less common intervention that did not provide students with any additional time used by one district: 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 academic recovery interventions targeted at subsets of students within each district, 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 Tier 1 interventions like math coaching for elementary teachers or a new literacy curriculum).

Across the eight school districts, we examine 12 interventions in math and 15 interventions in reading[2] across grades K-8.[3] We categorize the interventions into three distinct groups: 1) tutoring and small group instruction, 2) other supplemental instruction time programs, and 3) the "expert teacher" program. We primarily rely on observational methods to evaluate the effect of the programs on students' math and reading achievement. In a few instances where the program design and data allow for it, we use a Regression Discontinuity Design (RDD), but primarily we compare treated students with comparison students with similar observable characteristics, including prior achievement.

Consistent with prior studies of pandemic-era interventions, we find most of these 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). Across all tutoring, small group instruction, and other supplemental instruction time interventions, 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) but 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. Examining the expert teacher intervention, we find 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.

[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.

2. Background

2.1 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 (e.g., McCombs et al., 2017), computer-assisted learning (CAL) programs (e.g., Bettinger et al., 2022; Escueta et al., 2017), extended school days or years (e.g., Checkoway et al., 2013; Kraft, 2015; Kraft & Novicoff, 2024), double-dose courses in literacy (e.g., Arthur & Davis, 2016; Nomi, 2015; Özek, 2021) and grade retention (e.g., Jacob & Lefgren, 2009; Opper & Özek, 2024). 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. 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 three days per week.[5] However, the pre-pandemic research rarely included programs on the scale needed for COVID-recovery, raising questions

[4] For example, Jacob and Lefgren (2009) find that retaining low-performing eighth grade students increases the likelihood that these students later dropout 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.

about whether high-fidelity implementation is feasible (or necessary) for delivering promising programs at scale.

Tutoring programs in Nickow et al.'s (2024) meta-analysis, for example, had a median sample size of just 86 students for literacy and 173 for math—only some of whom received the treatment. Indeed, prior studies find a negative correlation between study sample-size and effect-size. Studies with smaller samples tend to have larger positive effects on student achievement than those with larger samples (Kohlmoos & Steinberg, 2024; Nickow et al., 2024). This negative association between program size and effectiveness is observed more broadly across education research (Kraft, 2020). Reduced researcher and/or provider oversight and increased variation in implementation when programs serve more students may explain some of this pattern (Hill & Erikson, 2019). The key point is that the pre-pandemic literature highlights promising programs but does not provide a clear road map on how to scale them up for COVID-recovery.

2.2 Implementing and Scaling Pandemic-Era Academic Recovery Interventions

During the second and third years of ESSER (2021-22 and 2022-23), scaling interventions with fidelity proved difficult. Districts 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 modified interventions to fit their resources and needs, sometimes diverging from central-office intentions or evidence-based practices, targeting the wrong students, or replacing core instructional time (Carbonari et al., 2024). The 2021-22 school year in particular included unprecedented problems for districts—including surges in infections due to new COVID variants, political polarization, and other challenges—that inevitably hindered districts' ability to implement interventions as intended.

Some adaptations may have been necessary to reach more students, but they risked compromising efficacy. For example, a district trying to scale up a tutoring program may have faced a trade-off between reducing tutoring hours per student and hiring less qualified tutors to expand the program. Emerging evidence suggests technology might help school districts address such dilemmas. Cortes et al. (2024), for example, conduct an RCT of the Chapter One literacy tutoring program serving 420 students (study sample = 818) across 49 kindergarten classrooms and find the program improves students' reading achievement by 0.11 SD. The technology-driven program costs $450/student and involves part-time tutors "pushing-in" to the classroom to provide short bursts (5-10 minutes) of instruction to individual students. Another recent RCT finds 0.23 SD gains in math scores for 2,060 9th grade students (study sample = 3,846) across two districts who took part in a daily tutoring program for 50 minutes per day (Bhatt et al., 2024). The program had pairs of students switch off daily between receiving in-person tutoring and computer-assisted learning (CAL), reducing costs (~$2,200 per student vs. ~$3,500 per student without CAL) and increasing its scalability.

Perhaps most relevant for the programs in the present study, Ready et al. (2024) find promising results for a large-scale virtual reading tutoring program serving students in grades 1-4. Their RCT estimates a 0.05 SD increase in reading achievement on NWEA's MAP Growth assessment for 959 treated students (study sample = 1,777) across six schools. The program offers 2-3 sessions per week for 30 minutes per session over 10 weeks (10-15 hours of treatment in total). Sessions occur during students' daily Learning Lab class period. In practice, however, about 20 percent of treated students completed the recommended dosage of at least 10 hours. On average, students received just 6.5 hours. Dosage varied widely across students and classrooms. Higher-performing students spent significantly more time using the program. Researchers also found a positive correlation between dosage and treatment effects. Despite some promising examples, the recent literature underscores the initial challenges school systems faced implementing pandemic-era interventions, especially at scale. However, it provides limited evidence on how targeted academic recovery efforts fared in 2022-23, when school districts returned to normal operations and were presumably poised to fully implement their interventions.

3. Methods

3.1 Sample

We examine academic recovery in the 2022-23 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 Road to Recovery research project.[6] We veil districts' names when reporting district-specific demographics or results to protect districts' 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 national averages.

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 academic recovery between districts by linking state test proficiency scores to the NAEP in 2022. In Table 2, we show the extent to which their test scores

[6] For more about the Road to Recovery project, including other research findings, see: https://caldercenter.org/covid-recovery.

in math and reading had recovered to pre-pandemic (i.e., spring 2019) levels as of spring 2023. Guilford County Schools's results show slightly smaller remaining gaps in math. Alexandria and Tulsa have large remaining gaps that range from -0.28 SD to -0.39 SD across math and reading. Based on average yearly pre-pandemic gains on interim assessments across grades 3-8, these larger declines are roughly equivalent to 80 to 110 percent of the gains students make in a typical year (Kuhfeld et al., 2024).

3.2 Data

This study uses NWEA's MAP Growth test scores, district-provided student-level data on demographics and eligibility for and participation in interventions, and information collected from interviews of district leaders about the design of interventions to analyze the impacts of each intervention on student achievement.

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, i.e. 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.[7] NWEA calculated these norms using MAP scores from a nationally representative sample of students from three pre-pandemic school years (i.e., 2016-16, 2017-18, and 2018-19). 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 identifiers that we link to school-level enrollment and demographic data from the 2020-21 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

[7] $z(Y_{\mathrm{igst}}) = (Y_{\mathrm{igt}} - \bar{Y}_{gt}) / SD(\mathrm{Y}_{\mathrm{gt}})$

lasted between 30 and 90 minutes. The interviews included questions about program content, program intensity, delivery mode, program providers, and student eligibility criteria.[8] Across the eight districts, we identified seven categories of academic recovery interventions: (1) tutoring programs, (2) small-group push-in and pull-out interventions, (3) after-school programs, (4) extended school years, (5) double-dose classes, (6) digital learning programs, and (7) assignment to an expert teacher.

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.[9]

We provide detailed information about the design characteristics (e.g., targeting criteria, delivery modality, provider type) of the (1) tutoring and small group instruction programs and (2) after-school, extended school year, double-dose, digital learning, and expert teacher programs respectively in Appendix Tables A1 and A2.

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

[8] 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.

[9] 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.

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 Appendix Table A2). These interventions also varied widely in terms of their total intended dosage, ranging from ~9 to ~124 hours over the course of the year.

3.3 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 a regression discontinuity design in a subset of cases). VAMs have often been used to estimate the impacts of schools on student outcomes (e.g., McEachin et al., 2016) as well as the impacts of interventions and policies on student achievement (e.g., Barry & Sass, 2022). We specify the following equation:

$$\begin{aligned} MAP_{ij,sub}^{Sp2023} = \beta_0 + \ \beta_1 Int_{i,sub} + \beta_2 Int_{i,sub}^{other} + \beta_3 Int_{i,\widetilde{sub}} + \delta MAP_{ij,sub}^{Fa2022} Grade_i \\ + \tau X_i Grade_i + \psi_{ij} + \varepsilon_{ij} \end{aligned} \tag{1}$$

where $MAP_{ij,sub}^{Sp2023}$ is the MAP Growth score for student *i* 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.[10] $Int_{i,sub}$ 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, 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.

We also include in our regressions the matrix $MAP_{ij,sub}^{Fa2022}$, 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 is restricted to students 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 $\widetilde{sub}$.[11] 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

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

[11] 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.

with grade level. Finally, ψ_{ij} denotes school-by-grade fixed effects and ε_{ij} represents idiosyncratic error. We estimate a linear model and calculate standard errors while clustering at the school-by-grade level (Abadie et al., 2022).[12]

To provide unbiased estimates of the effects of an intervention on student outcomes, VAMs must adequately control for all pretreatment variables that influence assignment to treatment and are related to students' outcomes. Researchers typically cannot rule out all potential sources of selection bias in a selection-on-observables design. For example, districts purposefully allowed teachers to use their subjective judgements of students' needs (in addition to measures such as test scores, grades, and attendance) to make program referrals. We typically do not have access to data on program referrals, only program participation. To address this concern, we conduct placebo tests that estimate 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: (1) that participating in a subject-specific intervention does not affect students' scores in the other subject and (2) 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.

[12] 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.

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

Regression Discontinuity

For certain interventions, we were also able to estimate treatment effects using a fuzzy regression discontinuity (RD) design. Researchers can apply the RD method when there is a clearly defined cutoff for an 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. 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, Todd & Van der Klaauw, 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 (2SLS) model:

$$Any_{sub,i} = \alpha_0 + \alpha_1 Elig_{sub,i} + \alpha_2 Score_{sub,i} + \alpha_3 \left(Score_{sub,i} * Elig_{sub,i}\right) + \gamma_i + e_i$$
$$MAP_{sub,i}^{Sp2023} = \beta_0 + \beta_1 \widehat{Any}_{sub,i} + \beta_2 Score_{sub,i} + \beta_3 \left(Score_{sub,i} * Elig_{sub,i}\right) + \gamma_i + u_i$$

Where Any$_{sub,i}$ is an indicator variable that takes the value 1 if student i received any treatment in subject sub during school year 2022-2023. Elig$_{sub,i}$ 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. Score$_{sub,i}$ 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 $MAP_{sub,i}^{Sp2023}$ and uses the estimated $\widehat{Any}_{sub,i,t}$ 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 0.5 SD, 0.75 SD and 1 SD of the running variable.

4. Results

4.1 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 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%. Across districts where we are able to identify targeted students, the percentage of targeted students who were subsequently treated varied widely, from 8% to 51%. For programs with available data on the number or percentage of students targeted for interventions, actual participation rates were consistently lower than the share of students program planners intended to serve.

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 Groups in reading). Programs varied even more dramatically in their minimum intended dosage per subject, ranging from 9 to 102 hours.[13] With the exception of four programs—District C Tutoring in math and reading, which has a notably wide range of intended dosage, and District B Tutoring in math and reading—the average hours actually attended by student participants were less than what was intended.

[13] 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. These estimates do not take into account the percent of sessions actually attended by students.

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. Participation rates and the extent to which targeted students were treated are more varied across these programs. For example, four of these programs (District C's Tutoring #2 in math and reading, and District H's After-School programming in math and reading) are universally available to all students. However, their take-up rates differ dramatically. Less than 1% of students utilized Tutoring #2 for math and only 1% of students utilized it for reading support in District C, while 24% of students in District H attended at least some of the After-School program. For the set of programs with some degree of targeting, participation rates sometimes fell short of the targeted share of students (e.g. District A Extended School Year in math and reading and District F Double-Dose in reading), though in some cases, participation exceeded the targeted rate (e.g. District D Digital Learning in math and reading). 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.

4.2 Intervention Impacts

Tables 5 and 6 report treatment effect estimates from value-added models of tutoring and small group interventions, and other supplemental time interventions, respectively. 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.[14] In columns 3-4 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 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.[15]

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 according to existing research (column

[14] Because the participation rates shown in Table 4-6 use the students included in the value-added analytic sample as the denominator, rates may differ from those shown in Table 3, which reports participation rates among all students in eligible grades in the district.

[15] 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 value-added models, 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.

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 SDs and in reading of 0.0089 SDs. 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 ($\beta = 0.033$, *p*>.05) or reading interventions ($\beta = 0.069$, *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 ($\beta = 0.218$, *p*<.001). 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. 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 Appendix Table A1): 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 9-102 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. However, it is also possible that teachers identified students—using measures not captured by our included prior test scores and covariates—for tutoring who were struggling specifically in math and not necessarily in reading, which would result in 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 result.

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 (*β* = 0.33, *p*<.001), and District B Tutoring (*β* = 0.23, *p*<.001). The placebo effect for the first is statistically indistinguishable from zero. For the second, the placebo effect is positive and significant (*β*=0.12, *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 hours 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 (*β* = -0.003, *p*>.05) or reading achievement (*β* = 0.0073, *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, *β* = 0.027), 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 affected by the treatment. The treatment 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 (β = 0.057, p<.001) and in reading (β = 0.108, 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 ($\beta_{placebo}$ = 0.023 vs. β_{main} = 0.108). 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, in Table 8, we report the results of the regression discontinuity analysis that we conduct for interventions in just two of the districts: District A Tutoring in both math and reading, and District D Digital Learning in both math and reading. The table shows estimates from the first and second stage of the fuzzy RD, across bandwidths of 0.5 SD and 1 SD of the running variable. We find that for all four interventions, the first stage is statistically strong (p<.001). 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 that there is no statistically significant discontinuity in test scores at the eligibility threshold. These findings are consistent with those from our value-added models of the same interventions, where we similarly were unable to detect significant impacts of the treatment.

For the handful of interventions where we do find significant evidence of positive impacts, 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.; however, they fell short of these estimates by only approximately 19% - 25%. Kraft (2020) reviews 750 randomized control trials (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.

5. Conclusion

In this study, we estimate the effects of academic COVID recovery interventions on student achievement during the 2022-23 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 1-2% 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 to 39 percent of students in grades 4-8—failed to produce significant improvements in student achievement, with one exception: 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. But, while such interventions may have benefited the students assigned to expert teachers, they would not be expected to raise achievement overall if they were offsetting relative losses for students assigned to non-expert teachers. For this reason, 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. But our findings suggest 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.[16]

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, 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: 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., 2024; Ready et al., 2024).

But 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; Fahle et al. 2024; Lewis & Kuhfeld, 2023), the need for action is urgent. States and the

[16] 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.

federal government must invest more in better understanding *how* to accelerate learning effectively before it is too late.

References

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. (2024). *A Scalable Approach to High-Impact Tutoring for Young Readers: Results of a Randomized Controlled Trial* (NBER Working Paper No. w32039). National Bureau of Economic Research.

Curriculum Associates (2023). *State of Student Learning in 2023.* https://cdn.bfldr.com/LS6J0F7/at/x8v8wp2c6j4s4wttsw2nwphb/ca-state-of-student-learningtechnical-report-2023.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: Cambridge, MA, USA*.

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

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: Cambridge, MA, USA*.

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: Cambridge, MA, USA*.

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.

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

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

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., & 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*.

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.

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

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.

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

Özek, U., & Mariano, L. T. (2023). Think Again: Is Grade Retention Bad for Kids?. *Thomas B. Fordham Institute*.

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

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.

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

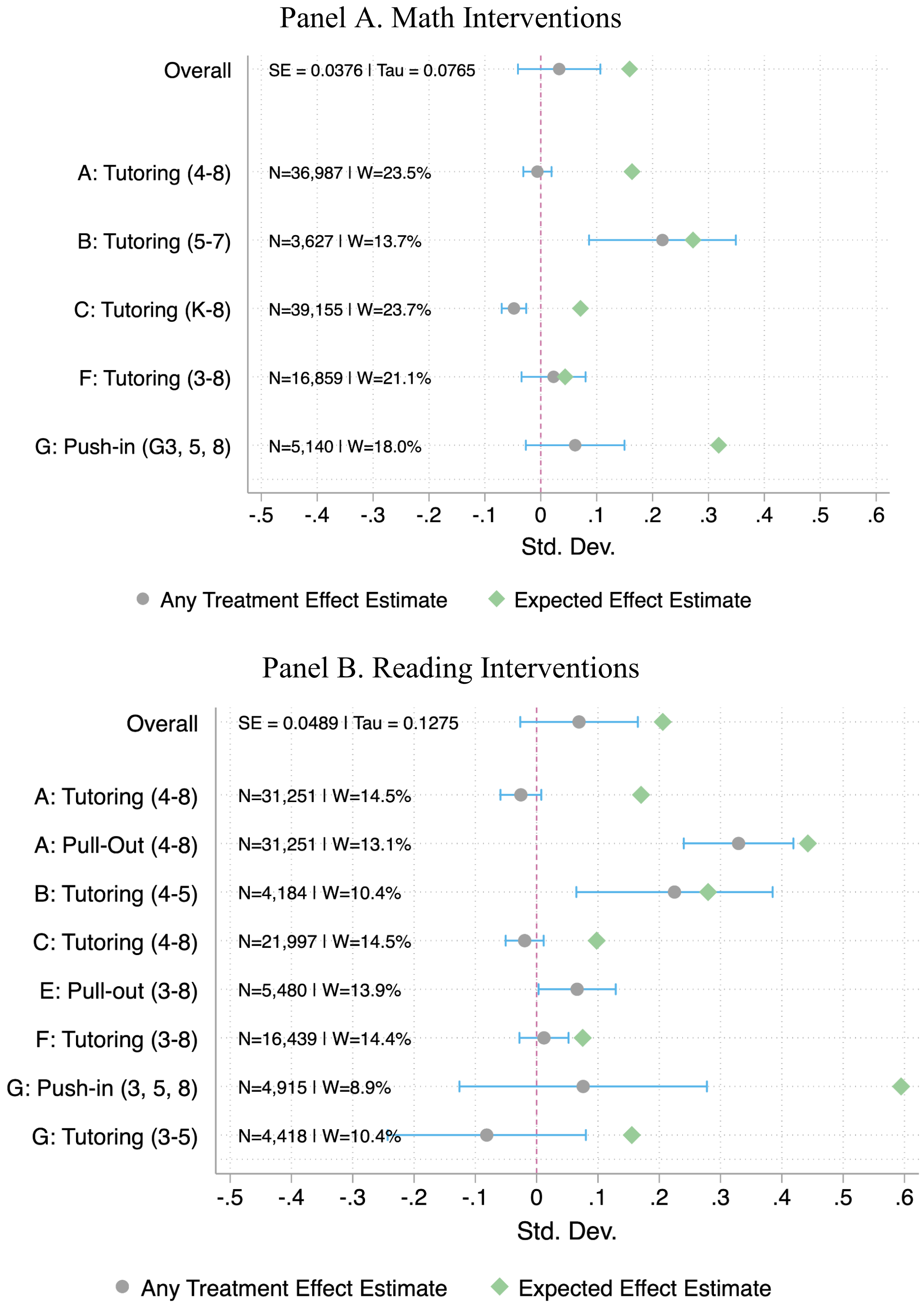
U.S. Department of Education. (2021). *Education in a pandemic: The disparate impacts of COVID-19 on America's students*. U.S. Department of Education, Office of Civil Rights

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

Figures and Tables

Figure 1. Estimated Treatment Effects of Tutoring and Small Group Instruction Interventions

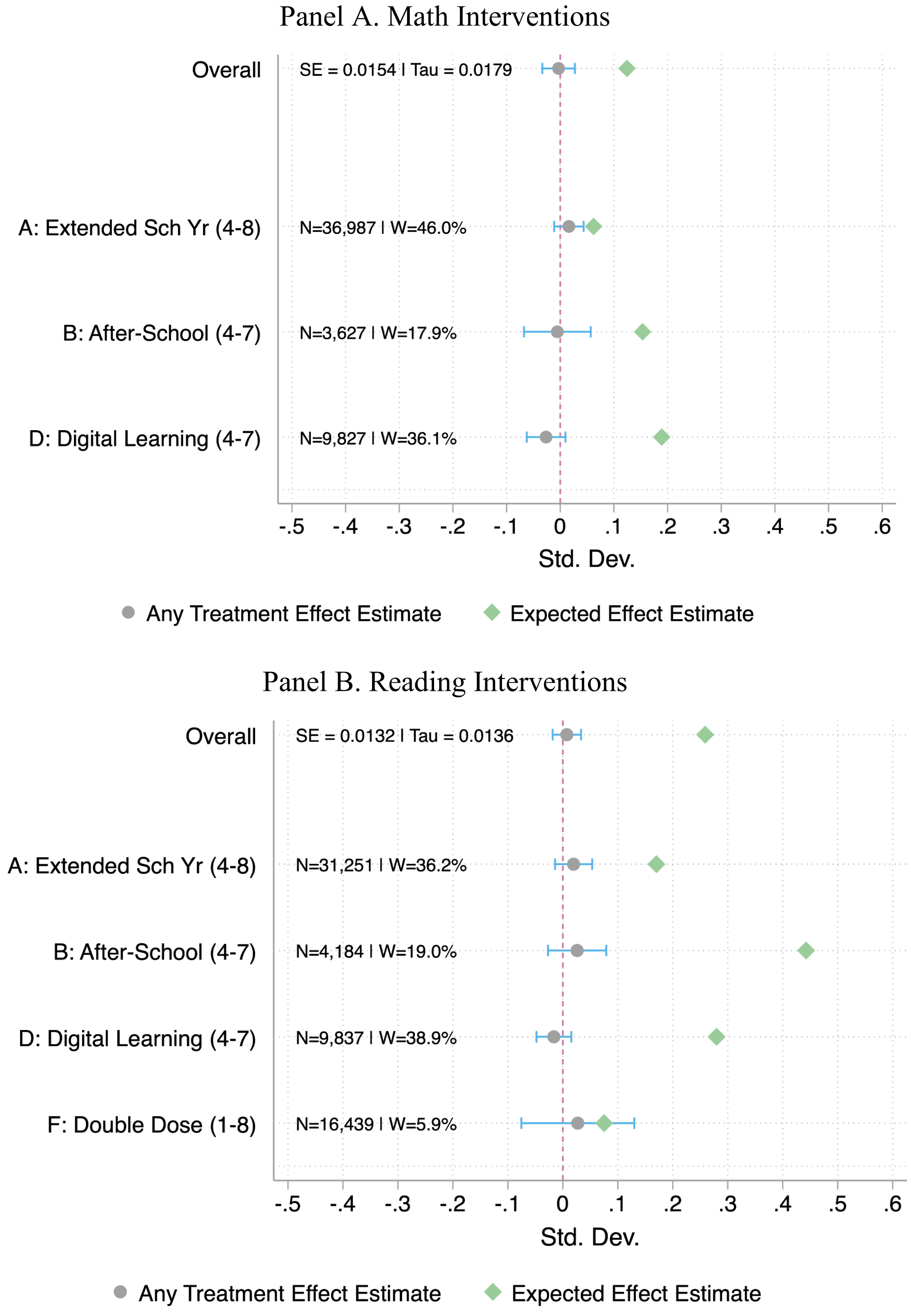


Panel A. Math Interventions

Panel B. Reading 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



Panel A. Math Interventions

Panel B. Reading Interventions

Table 1. Sample Demographics

	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. FRPL=free or reduced priced lunch. Data for the national sample and study district sample are from the Common Core of Data (CCD) collected by the National Center for Education Statistics during the 2022-23 school year. Statistics for the Nationwide NWEA sample are based on data from the 2019-20 CCD data collection, as reported in Isaacs et al. (2023).

Table 2. Estimated Achievement Loss and Recovery from Spring 2019 to 2023, Grades 3-8

	District	Spring 2019 (SDs)	Spring 2022 (SDs)	Spring 2023 (SDs)	Change from S19 to S22 (SDs)	Change from S22 to S23 (SDs)	Change from S19 to S23 (SDs)
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: Reading	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 standard deviation (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.

Table 3. Program Usage Across Sample Districts

Panel A: Math Interventions

	Tutoring	Small Group	After- School	Extended Calendar	Double-Dose	Digital Learning
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

	Tutoring	Small Group	After-School	Extended Calendar	Double-Dose	Digital Learning
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.

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	-	0%	-	30 - 60	38.73
District C: Tutoring (K-8)	39,155	24%	13%	21%	9 - 102	9.64
District C: Tutoring #2 (3-8)	21,997	100%	0%	0%	-	11.5 - 18
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	5.89
District G: Push-In Small Group(3, 5, 8)	5149	40%	7%	12%	35	27
B. Tutoring and Small Group Interventions - Reading						
District A: Tutoring (4-8)	56,407	32%	18%	51%	30	18.39
District A: Pull-Out Small Group (4-5)	22,713	34%	1%	4%	90	49.91
District B: Tutoring (4-5)	2,266	-	1%	-	30 - 60	31.36
District C: Tutoring (4-8)	21,997	22%	11%	17%	9 - 102	13.03
District C: Tutoring #2 (3-8)	21,997	100%	1%	1%	-	11.5-18
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	8.14
District G: Push-In Small Group (3, 5, 8)	4915	38%	8%	12%	35	31
District G: Tutoring (3-5)	4418	29%	4%	8%	24	16.33
C. Other Supplemental Time Interventions - Math						
District A: Extended School Year (4-8)	36,987	19%	13%	61%	18	7.99
District B: After-School (4-7)	4,149	-	8%	-	14	20.33
District D: Digital Learning (4-7)	11,702	29%	38%	85%	30	24.27
District H: After-School (3-8)	1,684	100%	24%	24%	-	4.5 - 21.75

D. Other Supplemental Time Interventions - Reading						
District A: Extended School Year (4-8)	56,407	19%	13%	61%	27	11.99
District B: After-School (4-8)	4,944	-	7%	-	14	10.89
District D: Digital Learning (4-7)	11,702	28%	34%	79%	30	27.96
District F: Double-Dose (6-8)	7,429	27%	4%	9%	124	56.62
District H: After-School (3-8)	1,684	100%	24%	24%	-	4.5 - 21.75
E. Expert Teachers						
Expert Teachers (4-8) - Math	36,987	-	26.40%	-	Full school year	Full school year
Expert Teachers (4-8) - Reading	56,407	-	19.40%	-	Full school year	Full school year

Note. Sample sizes shown reflect the unrestricted sample of students enrolled in grades eligible for the intervention. Cells left black either signify that there is 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.

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	101,768	12%	0.033	-	0.0015	-	21.5	0.16
			(0.038)	-	(0.0018)	-		
District A: Tutoring (4-8)	36,987	18%	-0.006	-0.025	-0.0003	-0.0011	22.1	0.16
			(0.013)	(0.020)	(0.0006)	(0.0009)		
District B: Tutoring (5-7)	3,627	1%	0.218**	-0.1941***	0.0059	-0.0052	37.0	0.27
			(0.067)	(0.026)	(0.0018)	(0.0007)		
District C: Tutoring (K-8)	39,155	13%	-0.048***	0.018	-0.0050	0.0019	9.6	0.07
			(0.011)	(0.015)	(0.0012)	(0.0015)		
District F: Tutoring (3-8)	16,859	3%	0.023	0.046	0.0038	0.0077	6.0	0.04
			(0.029)	(0.040)	0.0049)	(0.0066)		
District G: Push-In Support (3, 5, 8)	5140	3%	0.062	-0.070	0.0014	-0.0016	43.0	0.32
			(0.045)	(0.054)	(0.0010)	(0.0013)		
			B. Reading Interventions					
Overall	119,935	11%	0.069	-	0.0030	-	23.1	0.21
			(0.049)	-	(0.0021)	-		
District A: Tutoring (4-8)	31,251	22%	-0.026	-0.006	-0.0013	-0.0003	19.1	0.17
			(0.017)	(0.014)	(0.0009)	(0.0007)		
District A: Pull-Out Small Group (4-5)	31,251	2%	0.3294***	0.030	0.0066	0.0006	49.7	0.44
			(0.046)	(0.038)	(0.0009)	(0.0008)		
District B: Tutoring (4-5)	4,184	1%	0.2249**	0.1235*	0.0075	0.0041	31.4	0.28
			(0.082)	(0.058)	(0.0027)	(0.0019)		
District C: Tutoring (4-8)	21,997	11%	-0.020	0.014	-0.0018	0.0013	11.0	0.10
			(0.016)	(0.013)	(0.0014)	(0.0012)		
District E: Pull-Out Small Group (3-8)	5,480	12%	0.066*	0.092**	-	-	-	-
			(0.032)	(0.032)				
District F: Tutoring (3-8)	16,439	13%	0.012	-0.015	0.0015	-0.0019	8.4	0.08
			(0.021)	(0.025)	(0.0026)	(0.0032)		
District G: Push-in Small Group (3, 5, 8)	4,915	4%	0.076	-0.114*	0.0011	-0.0017	66.8	0.59
			(0.103)	(0.052)	(0.0015)	(0.0008)		
District G: Tutoring (3-5)	4,418	2%	-0.081	-0.080	-0.0047	-0.0046	17.4	0.16
			(0.083)	(0.075)	(0.0047)	(0.0043)		

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

Note. Main effect point estimates show the average effect of receiving any amount of math (or reading) intervention during 2022-23 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 value-added models 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, i.e. 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 restricted maximum likelihood (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.

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%	-0.003	-	-0.0002	-	16.81	0.1244
			(0.0154)	-	(0.0009)	-		
District A: Extended School Year (4-8)	36,987	14%	0.0162	-0.0083	0.0019	-0.0010	8.4	0.06
			(0.0140)	(0.0183)	(0.0017)	(0.0022)		
District B: After-School (4-7)	3,627	8%	-0.0054	-	-0.0001	-	20.7	0.15
			(0.0317)	-	(0.0009)	-		
District D: Digital Learning (4-7)	9,827	39%	-0.0263	-0.0142	-0.00103	-0.00056	25.6	0.19
			(0.0184)	(0.0182)	(0.0007)	(0.0007)		
			B. Reading Interventions					
Overall	61,711	15%	0.0073	-	0.0003	-	29.09	0.2589
			(0.0132)	-	(0.0004)	-		
District A: Extended School Year (4-8)	31,251	13%	0.0196	0.0173	0.0016	0.0014	12.0	0.11
			(0.0172)	(0.0149)	(0.0014)	(0.0012)		
District B: After-School (4-8)	4,184	7%	0.0260	-	0.0007	-	11.2	0.10
			(0.0270)	-	(0.0007)	-		
District D: Digital Learning (4-7)	9,837	35%	-0.01625	0.0017	-0.00058	0.0001	28.2	0.25
			(0.0162)	(0.0174)	(0.0006)	(0.0006)		
District F: ELA Double Dose (1-8)	16,439	1%	0.0272	-0.0588	0.0004	-0.0008	69.5	0.62
			(0.0524)	(0.0373)	(0.0008)	(0.0005)		

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

Note. Main effect point estimates show the average effect of receiving any amount of math (or reading) intervention during 2022-23 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 value-added models 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, i.e. 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 restricted maximum likelihood (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.

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***	0.0005
			(0.0135)	(0.0138)
Expert Teachers in ELA (4-8)	31,251	20%	0.1083***	0.0230
			(0.0140)	(0.0116)

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

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 2022-23 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 value-added models 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.

Table 8. Estimated Treatment Effects from Regression Discontinuity Models

	Bandwidth	Model	N	Point Estimate	SE
District A: Tutoring					
Math	0.5 SD	First Stage	9,154	0.7715***	0.0114
	0.5 SD	Second Stage	9,154	0.0081	0.0433
	1 SD	First Stage	16,410	0.7743***	0.0088
	1 SD	Second Stage	16,410	-0.0314	0.0270
Reading	0.5 SD	First Stage	7,212	0.8300***	0.0124
	0.5 SD	Second Stage	7,212	-0.0098	0.0533
	1 SD	First Stage	14,028	0.8271***	0.0095
	1 SD	Second Stage	14,028	0.0166	0.0381
District D: Digital Learning					
Math	0.5 SD	First Stage	2,989	0.3329***	0.0384
	0.5 SD	Second Stage	2,989	0.1272	0.1342
	1 SD	First Stage	5,439	0.3662***	0.0088
	1 SD	Second Stage	5,439	0.0233	0.0877
Reading	0.5 SD	First Stage	2,479	0.4647***	0.0511
	0.5 SD	Second Stage	2,479	0.0653	0.1453
	1 SD	First Stage	5,049	0.4595***	0.0397
	1 SD	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.

Appendix

Appendix Table A1. Designs of Tutoring and Push-in or Pull-out Small Group Instruction Programs (K-8)

District	Subject	Grades (K-8)	Eligible schools	Student eligibility criteria	Modality	During school?	Counter-factual	Provider-to-student ratio	Provider type	Intended frequency and duration	Intended session length	Total intended instruction time per subj per year
A	Math, reading, science	4-8	All schools	Scoring below state test threshold	Mostly in-person, but some virtual	Mostly during school, but some after school	Elective classes or out-of-school time	1:10 max, targeted 1:3 to 1:6	District teachers and staff, retired teachers, and college students	Varied by vendor	30-90 min	30 hrs
A	Reading	1-5	~13% of schools with lowest test scores and highest percentage of Black students	Scoring below NWEA MAP threshold, not SPED, teacher recs	In-person	During school	Varies by campus, not supposed to pull students out during core reading instruction	1:4 to 1:6	Certified district teachers with specialization in reading	4 days/wk, full school year	45 min/day	~108 hours
B	Math, reading	4-7	~20% of schools	Scoring in a range of NWEA MAP scores	Virtual	After school	Out-of-school time	1:2 or 1:3	Undergraduate and graduate students	3 days/wk, ~10-20 weeks	60 min	30-60 hrs
C (#1)	Math, reading, science, social studies	K-8	All schools	Scoring below NWEA MAP threshold, teacher recs	In-person	Mostly during school, but some after school	Intervention block (small group instruction, digital learning programs)	1:1 to 1:5	Certified teachers, community members, graduate students, high school students	2-3 days/wk, 9-34 weeks	30-60 min	9-102 hrs

C (#2)	Math, reading, science, social studies	3-8	All schools	All students, opt-in	Virtual	After school	Out-of-school time	1:1	Certified district teachers	N/A, on-demand	varies	N/A
E	ELA	K-3, 6-8	All schools	Lowest scoring on district literacy tests	In-person	During school	Small group instruction during reading intervention block	Grades K-2: 1:6 Grades 6-8: ~1:20	Certified teacher and, for grades K-2, a teaching assistant	5 days/wk, full school year	Grades K-2: 30 min/day Grades 6-8: 40-45 min/day	Grades K-2: ~90 hours Grades 6-8: ~120-134 hours
F	Mostly reading, some math and social studies	3-8	All schools	Scoring in a range of NWEA MAP scores, teacher recs	Virtual	Mostly during school, but some after school	Intervention block (small group instruction, digital learning programs)	1:1	Hired and trained by vendor; required to have a BA and 2 years teaching or tutoring experience; tutor and student pair could vary each session	Varied by classroom (max ~2 days/wk), recommended 40 min/wk for the whole school year	20-60 min	~12-36 hrs
G	Reading	3-5	~50% of schools based on % all students and # Black students scoring below NWEA MAP threshold	Scoring in a range of NWEA MAP scores, teacher recs	In-person	After school	Out-of-school time	1:3 max	Certified district teachers and staff	3 days/wk, 8 or 16 weeks	30-45 min	12-36 hrs

G	Math and reading	3, 5, 8	~50% of schools based on # of Black and Native students scoring below state test threshold	Black, Native, and Black or Native multiracial students	In-person	During school	Receiving core instruction from the classroom teacher without support from an additional instructor in a small group	1:6 to 1:7	Certified district staff with specialization in reading or math	4 days/wk, full school year	15 min/day	~36 hours

Appendix Table A2. Designs of Other Supplemental Time Interventions (K-8)

District	Program type	Subject	Grades (K-8)	Eligible schools	Student eligibility criteria	Modality	During school?	Counter-factual	Provider -to-student ratio	Provider type	Intended frequency and duration	Intended session length	Total intended instruction time per subj per year
A	Extended School Year	Math and ELA	K-8	~28% schools	Test scores, other risk factors, opt-in	In-person	During school vacation	Out-of-school time (school vacation)	1:15 max	Certified campus teachers	18 days over 5 weeks (with 3-4 days per week), but students could choose to attend any 1-5 weeks	Typical school day, ~7.5hrs. 1.5 hrs/day of reading, 1 hr/day of math	Math: 3-18 hours Reading: 4.5-27 hours
B	After-school	Math, reading	K-8	Mostly Title I schools	State test scores, race, and special program (e.g., SPED, ELL, FRL) status	In-person	After school	Out-of-school time	1:10 to 1:20 max	Certified and non-certified district staff	3 days/wk min, academic instruction occurs ~18 weeks/yr	~2.5 hrs/day, with ~30 min/day of instruction time	~14 hours

D	Digital learning	Math and reading	K-8	All schools	Scoring below state test or NWEA MAP threshold	In-person	During school	Intervention block (tier 2 students) or enrichment (tier 1 students)	1:6 max	Certified district teachers and interventionists	Daily until student has received at least 30 hours (but can receive more)	Varies	At least 30 hours
F	Double-dose class	ELA	6-8	~20% middle schools	Scoring in a range of NWEA MAP scores	In-person	During school	Elective courses	1:15 at one site, 1:25-30 at the other site	Certified district teachers	5 days/wk, full school year	45 min/day	~124 hours
H	After-school	Math and ELA, enrichment	3-8	All schools	Opt-in	In-person	After school	Out-of-school time	1:10 to 1:15	Certified and non-certified district staff	Grades 3-5: 2 days/wk, ~24 weeks Grades 6-8: 1-4 days/wk, ~24 weeks	Grades 3-5: ~1.5 hrs/day, including 30 min/day of total (math and ELA) instruction Grades 6-8: 1 hr/day, including 45-60 min of total (math and ELA) instruction	Grades 3-5: ~12 hours Grades 6-8: ~9-48 hours

–	Expert teachers	Math and ELA	4-8	All schools	All students	In-person	During school	Non-expert designated teacher	Typical class size	Teachers designated by the district as expert teachers based on observations, student growth, and National Board certification	N/A, replaces regular instructional time	N/A, replaces regular instructional time	N/A, replaces regular instructional time

Note. District D’s digital learning program intervention may also have included some small group instruction, but the data we received primarily captured time spent on digital learning programs. We do not disclose the district that implemented the expert teachers intervention to preserve district anonymity.