The Impact and Implementation of Academic Interventions During COVID: Evidence from the Road to Recovery Project

Maria V. Carbonari
Miles Davison
Michael DeArmond
Daniel Dewey
Elise Dizon-Ross
Dan Goldhaber
Ayesha K. Hashim
Thomas J. Kane
Andrew McEachin
Atsuko Muroga
Emily Morton
Tyler Patterson
Douglas O. Staiger

June 2024

WORKING PAPER No. 275-0624-2

CALDER
National Center for Analysis of Longitudinal Data in Education Research

AIR®

The Impact and Implementation of Academic Interventions During COVID: Evidence from the Road to Recovery Project

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 / CEDR

Emily Morton
American Institutes for Research / CALDER

Miles Davison
NWEA

Ayesha K. Hashim
NWEA

Andrew McEachin
NWEA

Tyler Patterson
University of Chicago

Douglas O. Staiger
Dartmouth College

Contents

Acknowledgments

This research was supported by the Carnegie Corporation of New York, the Walton Family Foundation, Kenneth C. Griffin, the AIR Equity Initiative, and an anonymous foundation. We could not have drafted this report without the district leaders who generously gave their time and attention to us during a challenging school year. We are grateful to Anna McDonald, Ian Callen, Andrew Camp, and Andrew Diemer for their assistance with various aspects of this work.

We are grateful to Anna McDonald, Ian Callen, Andrew Camp, and Andrew Diemer for their assistance with various aspects of this work.

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

The Impact and Implementation of Academic Interventions During COVID: Evidence from the Road to Recovery Project

Maria V. Carbonari, Miles Davison, Michael DeArmond, Daniel Dewey, Elise Dizon-Ross, Dan Goldhaber, Ayesha K. Hashim, Thomas J. Kane, Andrew McEachin, Atsuko Muroga, Emily Morton, Tyler Patterson, and Douglas O. Staiger

CALDER Working Paper No. 275-0624-2
June 2024

Abstract

In this paper we examine academic recovery in 12 mid- to large-sized school districts across 10 states during the 2021–22 school year. Our findings highlight the challenges that recovery efforts faced during the 2021–22 school year. Although, on average, math and reading test score gains during the school year reached the pace of pre-pandemic school years, they were not accelerated beyond that pace. This is not surprising given that we found that districts struggled to implement recovery programs at the scale they had planned. In the districts where we had detailed data on student participation in academic interventions, we found that recovery efforts often fell short of original expectations for program scale, intensity of treatment, and impact. Interviews with a subsample of district leaders revealed several implementation challenges, including difficulty engaging targeted students consistently across schools, issues with staffing and limitations to staff capacity, challenges with scheduling, and limited engagement of parents as partners in recovery initiatives. Our findings on the pace and trajectory of recovery and the challenges of implementing recovery initiatives raise important questions about the scale of district recovery efforts.

1. Introduction

Pandemic-era disruptions to schooling have resulted in academic setbacks for many students in the US. The pandemic's negative impact on learning is reflected in a range of assessments, from the National Assessment of Educational Progress (NAEP) (U.S. Department of Education, 2022a; 2020b) to NWEA's MAP Growth tests (Kuhfeld & Lewis, 2022; 2023) and Curriculum Associates' i-Ready assessments (Curriculum Associates, 2020). Besides generally harming academic progress, pandemic disruptions have worsened prepandemic inequities by disproportionately impacting students with lower test scores and students from historically marginalized groups (Dorn et al., 2021; Education Policy Innovation Collaborative, 2021; Lewis et al., 2021).[1]

School districts nationwide have responded with a range of interventions to help students catch-up academically, aided by $190 billion from the American Rescue Plan's Elementary and Secondary School Emergency Relief (ESSER) fund. Popular interventions include teaching students in small groups, offering one-on-one tutoring, adding classes before and after school, and adding instructional minutes to the school day (Diliberti & Schwartz, 2022). The stakes surrounding districts' academic recovery efforts are high. Hanushek (2023), for example, estimates that students who fell behind during the pandemic could see their lifetime earnings fall by 2-9 percent, and states could see their GDPs decrease by 3.5 percent, on average. Using changes in earnings in states with prior increases on the NAEP, Doty et al. (2022) estimate smaller, but still sizable impacts on earnings of 1.6 percent. Beneath these averages, the

[1] Besides its negative impact on student learning, the pandemic has also negatively impacted students' social and emotional well-being (Bradshaw et al., 2023; Hamilton & Gross, 2021). Early in the pandemic, for example, survey research revealed declines in reported mental health (Patrick et al., 2020). Subsequent surveys suggest mental health challenges continued for students in the wake of COVID, especially for LGBTQ students and girls (Jones et al., 2022).

pandemic's disparate impact raises urgent concerns about equity and earnings inequality. As the U.S. Department of Education's Office of Civil Rights noted in a 2021 report, students "who went into the pandemic with the fewest opportunities are at risk of leaving with even less" (U.S. Department of Education, 2021, p. 51).

In this paper, we use multiple data sources to assess academic recovery efforts in four school districts. The districts are part of an ongoing collaboration between districts and researchers at the American Institutes for Research, Harvard University, and NWEA. Our analysis of participation and achievement test data suggests that the districts' interventions during the 2021-2022 school year failed to reach the intended number of students, and few had statistically or practically significant effects on student math and reading test scores through spring 2022. Interviews with district leaders in three of the four districts (with interventions we assess) highlight a host of implementation challenges districts faced during the 2021-2022 school year, including challenges reaching target populations, staffing interventions, scheduling interventions, accommodating existing policies, and building adequate central office capacity.

Taken together, these results are important not only for districts' near-term recovery efforts but also how districts can respond to future recovery efforts coming out of periods of disrupted learning (e.g. natural disasters) (Opper et al., 2023). Indeed, we estimate that—even if programs had yielded the same large effects associated with high-dosage tutoring programs in the prepandemic literature (Nickow et al., 2024)—the planned scale (i.e., participation rate and dosage) of the four districts' recovery interventions for the 2021-22 school year would not have been enough to address the full scale of their students' academic recovery needs. If K-12 systems are not able to improve and expand their efforts to help students catch up, pandemic losses could have long-term implications for equity and opportunity in the US.

2. Background

COVID-19's negative impact on academic achievement in K-12 schools has been well documented. Two years after the pandemic upended schools nationwide, results from the NAEP's 2022 long-term trend assessments marked the nation's largest drop in reading scores since 1990, and the first ever drop in mathematics scores (U.S. Department of Education, 2022a); these results were soon followed by historic drops in the main NAEP assessments in reading and mathematics (U.S. Department of Education, 2022b). To help put the losses in perspective, Fahle et al. (2023) estimate the magnitude of the average decline is roughly equivalent to half a grade level in math and almost a third of a grade level in reading.[2,3] But the pandemic's effects were not uniform. Across assessments and studies, the academic losses have generally been worse in math than reading, worse for students who spent more of the 2020-21 school year in a remote or hybrid learning environment, and worse for students living in low-income households and those from historically marginalized groups (Fahle et al., 2023; Goldhaber et a., 2022a; Goldhaber et al., 2022b; Lewis et al., 2021; West & Lake, 2021). Among districts that operated remotely for most of the 2020-21 school year, for example, students in districts serving a high percentage of minority students were the equivalent of 0.8 grade levels behind their prepandemic scores in math, while students in low minority districts were about 0.5 grade levels behind (Fahle et al., 2023). To further contextualize the scale of these losses, we note the magnitude of the test score declines in math was similar to (if not a bit larger than) that of the historically large declines

[2] Compared to the 2019 scores, the 2022 NAEP scores were 0.20 standard deviations lower in 8th grade math, 0.08 standard deviations lower in 8th grade reading, 0.15 standard deviations lower in 4th grade math, and 0.08 standard deviations lower in 4th grade reading.

[3] Translations of effect size declines to equivalent declines in grade levels or weeks of learning are useful for contextualizing impacts. However, they should be interpreted with caution, as rough (rather than precise) approximations of impacts, due to the statistical assumptions required for the calculation (see Baird & Pane, 2019; Kuhfeld & Soland, 2021).

experienced by evacuees of Hurricane Katrina, one of the worst natural disasters in U.S. history (Sacerdote, 2012).

During the 2021-22 school year—the time-period covered by this study and, for many districts, the first school year "in-person" since the pandemic—the school-year pace of academic growth mostly returned to prepandemic rates (Kuhfeld & Lewis, 2022). But to close the gap between pre- and post-pandemic test scores, the pace of academic growth needs to be faster than "normal." During the 2022-2023 school year, the pace of learning was not significantly faster; in fact, it may have been slightly slower. As a result, the *average* student in grades 3-8 needs an extra four to five months of instruction to reach prepandemic achievement levels in math and reading (Lewis & Kuhfeld, 2023). And for historically marginalized students disproportionately impacted by the pandemic, the timeline for academic recovery is even longer. Just to return to prepandemic levels of inequality, students attending high-poverty schools are estimated to need the equivalent of an *additional* month of schooling relative to students attending low-poverty schools (Isaacs, Kuhfeld, & Lewis, 2023).

2.1 District Recovery Efforts

Prepandemic research suggests some of the academic interventions that districts are using to deal with pandemic losses—like tutoring—have the potential to accelerate student learning (Nickow et al., 2024). At the same time, research unsurprisingly suggests that the relationship between an intervention and academic outcomes is mediated by the intervention's design and implementation (e.g., Lynch et al., 2022; McEachin et al., 2018; Nickow et al., 2024). The promising results on tutoring, for example, rely on "high dose" designs that provide tutoring to small groups multiple times per week during the school day throughout the school year (Harris, 2009; Nickow et al., 2024). Meanwhile, the delivery of supports like tutoring is influenced by

broader implementation issues, including the supply of providers, leadership commitment, coordination dynamics, and scheduling logistics (White et al., 2023). Stepping back, a broader literature underscores how front-line implementation is further complicated by the institutional context surrounding schools, as multiple actors—those who deliver interventions but also school leaders, central office staff, superintendents, school boards, and other policymakers—influence which intervention options are considered and the level of resources available to support them (Meier et al., 2004; Sandford & Moulton, 2015). Despite the growing empirical literature on the negative consequence of the pandemic and the stakes surrounding recovery, little is known beyond a few cases about the extent to which specific district responses are helping students rebound (Barry & Sass, 2022; Cortes et al., 2023).

In the next section, we describe our study methods, including our sample, data, and analytic approach. Then we review our findings on impact and implementation and end with a discussion of the results and their implications.

3. Methods

3.1 Sample

This study investigates academic recovery efforts in a sample of four districts to understand whether and how districts' responses provided students opportunities to catch up to prepandemic levels of achievement. These large, urban school districts were recruited[4] during the summer of 2021 to be part of the Road to COVID Recovery (R2R) research project.[5]

[4] To participate in R2R, we required that the district tested with NWEA in 2021-22, had plans to launch or expand an academic recovery program in 2021-22 that provided additional instructional time to students, and collected student-level data on participation in this program. Researchers provided districts with information about the opportunity to participate through the districts' existing contacts at NWEA. Participation was entirely voluntary and did not include any compensation.

[5] These four districts were the only districts of the R2R districts in the larger project during the 2021-22 school year that provided sufficient data to estimate program impacts before the requested deadline. Of the remaining districts, at least four had not implemented academic interventions at a scale and/or did not collect data that would enable the research team to conduct a meaningful impact analysis.

Together, the districts enroll over 340,000 students across three states. As shown in Table 1, the districts serve higher proportions of students of color and students attending high-poverty schools compared to national averages.

3.2 Data

We use a combination of quantitative and qualitative data to examine academic interventions during the 2021-2022 school year. The study's main conclusions about academic recovery and impact rely on the quantitative data.

Quantitative Data

The quantitative data for our study come from student achievement test scores on the NWEA Measures of Academic Progress (MAP) Growth math and reading assessments in grades 3–8. The MAP Growth test has several advantages for measuring academic recovery. First, the tests are administered in fall, winter, and spring, allowing us to gauge changes in achievement during the school year. This is important for assessing pandemic recovery interventions because some did not launch until the second half of the year. Second, the tests are computer adaptive (i.e., item difficulty increases or decreases in response to performance). Adaptive tests like MAP Growth are more precise at the high and low ends of the achievement distribution, which is useful for assessing pandemic recovery given the disproportionate effects of the pandemic on students who were already struggling academically (Kingsbury et al., 2014). Third, its items are linked to a common vertical scale that allows us to compare achievement and growth within and across districts.

The study districts also provided detailed student-level eligibility and participation data on their academic recovery interventions[6] that allowed us to examine how many and which

[6] Interventions for which student-level participation data were not available were not possible to include in the analysis.

students participated, how long they participated (e.g., days) and at what level of intensity (e.g., hours per day), and the impact of the intervention on math and reading achievement. Per our agreements with the districts, we veil their names when reporting our results and are purposely ambiguous when describing interventions to protect their anonymity. Appendix Tables A1 and A2 respectively display the math and reading standardized MAP Growth scores for the sample for each intervention by treatment status and term.

Qualitative Data

To identify the academic interventions for the study, we collected detailed programmatic data from documents and interviews on recovery efforts in each district.[7] Prior to data collection, we defined academic recovery interventions as programs that (a) were new or had expanded since the pandemic, (b) were supported by ESSER funds, and/or (c) provided targeted students with additional learning time beyond what was offered during standard instruction. Over the course of the school year, we interviewed small groups of district staff and program leaders selected by each district for their knowledge of the district's academic covid recovery interventions, resulting in a dataset of eight interviews across 22 total staff members. The identified interventions fell into five categories: (a) tutoring programs, (b) small-group push-in and pull-out interventions (c) out-of-school-time programs (d) virtual learning programs, and (e) extended school-year calendars. For the purposes of this study, we collapse tutoring programs

[7] We collected qualitative data about the districts' academic recovery interventions throughout the school year via two waves of semi-structured interviews. The interview protocols focused on central office intervention leaders and lasted between 60 and 90 minutes. We conducted interviews in fall 2021 and spring 2022 to develop detailed program descriptions. A notetaker on the research team shared his or her notes (templates for these notes are publicly available at covidrecovery.us with participants in real time so participants could check our descriptions for accuracy during the interview (to clarify outstanding questions, we also followed up via email and reviewed any documentation shared by district leaders). These interviews gave us detailed information about program design choices. Design choices included program type (e.g., tutoring, virtual learning), program content subject area (e.g., math or reading), program intensity (sessions per week), program dosage (minutes per session), program duration (days or weeks per year), delivery mode (e.g., virtual or in-person), provider types (e.g., teachers, community members), and student eligibility criteria.

and small-group pull-out interventions into one category because of the similarities in the design of the two types of interventions. The interventions implemented in each of the four districts and details on their designs are respectively displayed in Table 2 and Appendix B.

Besides interviewing district staff about intervention designs, we conducted additional interviews with district-level program leaders[8] in three of the four districts (the fourth district declined to participate). In these additional interviews, we used the results of the impact analysis as a jumping off point for probing the leaders about implementation factors that might explain the results. These interviews took place in the summer of 2022, lasted between 60 and 90 minutes, and covered a range of implementation issues, including intervention participation, perceptions about what was working and not working, challenges and barriers, and the intervention's future. Table 3 describes the number of administrators interviewed for each district and the interventions covered in the supplemental interviews.

4. Analysis

4.1 Impact Analysis

We estimate the impact of each recovery intervention using a value-added framework that controls for observable pretreatment student characteristics, as well as pretreatment test scores. This approach has been used to understand the impact of schools on student outcomes in general (e.g., McEachin et al., 2016), as well as to evaluate the impact of educational programs and policies on students' achievement (Barry & Sass, 2022).

Value-added methods can provide unbiased estimates of intervention impacts if students' assignment to treatment is as good as random after conditioning for observable pretreatment characteristics. While "pretreatment" might typically be interpreted as the start of the school year

[8] These leaders' titles varied across districts but included titles such as PK-12 Director of Math, Deputy Chief of Teaching and Learning, Multi-Tiered Systems of Support (MTSS) Director, and Tutoring Director.

(fall 2021) and earlier, in several of our participating districts, we saw evidence that student assignment to treatment was additionally based on measures of academic progress that became available *during* the school year. Specifically, second semester participation was related to students' winter 2021-22 MAP Growth assessment scores—even after controlling for earlier pretreatment test scores (i.e. from fall 2021 and the prior spring 2021). If students struggling academically mid-school year were more likely to be assigned to second semester treatment, then our impact estimates would be negatively biased unless we condition on mid-year test scores in addition to earlier pretreatment characteristics.[9,10]

To account for this scenario of mid-year treatment assignment, we therefore estimated the following semester-level model:

$$MAP_{igjts} = \alpha_0 + \alpha_1 Treatment_{igt} + \alpha_2 Eligible_{igts} + priorMAP_{igts}\gamma + X_{igt}\theta + \delta_{jgt} + \epsilon_{igts}$$

Here, MAP_{igjts} is the end-of-term MAP Growth score for student *i* in grade *g* at school *j* in semester *t* and subject *s.* We standardize these scores at the subject and grade level using pre-pandemic NWEA national norms[11], so that the outcome can be interpreted as MAP Growth performance relative to the national distribution of students prior to the pandemic. $Treatment_{igt}$ is a vector of binary indicators of treatment receipt for all recovery interventions available in the district in semester *t*.[12] We include measures of students' participation in any available intervention in order to isolate the effect of participation *only* for the treatment in question, as it

[9] Note that our estimate would still be biased if program participation were associated with other unobservable student characteristics, such as socio-emotional challenges, home situations, or course grades.

[10] Prior studies of the validity of value-added estimates (Kane and Staiger (2008), Chetty, Friedman, and Rockoff (2014), Kane, McCaffrey, Miller and Staiger (2013)) have suggested that value added measures generate unbiased predictions of impacts of teachers or schools. However, in those cases, students are assigned to a teacher or school at the beginning of the year, not mid-year.

[11] See Thum and Kuhfeld (2020) for more detail on the calculation of prepandemic norms.

[12] See Appendix C for discussion of using continuous measures of treatment receipt.

is possible in many cases for students to participate in multiple interventions simultaneously. For some recovery interventions, students were supposed to be eligible to participate if they scored below a certain level on a previous MAP Growth assessment or other standardized test.[13] In those cases, $Eligible_{igts}$ is a binary indicator for whether student i met the intervention eligibility requirements, interacted with grade level. $priorMAP_{igts}$ is a matrix with a cubic function of the start-of-term MAP Growth score in the same subject, as well as a cubic function of the same-subject score from one term prior, interacted with grade level and the term in which the treatment occurred (spring or fall). X_{igt} is a vector of baseline student characteristics, including race and ethnicity, gender, special education status, disability status, free or reduced-price lunch (FRPL) eligibility, and English Language Learner (ELL) status, as well as the start-of-term MAP Growth score in the other tested subject and the instructional week in which the end-of-term MAP Growth assessment was taken. δ_{jgt} contains school-grade-semester-level fixed effects.

The coefficient of interest from this model is α_1 for the treatment in question, which can be interpreted as the difference in MAP performance at the end of the semester (in either math or reading) between observably similar treatment participants and non-participants, within the same grade and the same school, holding constant their prior MAP performance and participation in any simultaneously offered recovery programs.

In one district, MAP Growth testing rates were notably low in spring 2022, with roughly 50 percent of tested grades not taking the assessment in that final term. As a result, in that

[13] Though some interventions were designed to target students who scored below certain thresholds, we do not see that districts adhered to these criteria when assigning students to treatment in practice. Appendix Tables A3 and A4 provide details about the intersection of intended eligibility (where applicable) and actual treatment for each program. As described further in Appendix C, we explored the potential for regression discontinuity analyses for each of these interventions, but found the first stage was too weak to warrant the analysis in each case.

district, we estimated the impact of first semester treatment participation only, using fall 2021 scores as the baseline achievement measure and winter 2022 results as the outcome.

Generally, the analytic sample for each district is limited to those students who had MAP Growth assessment scores from the start and end of the term in which the treatment took place (e.g., fall 2021 and winter 2021-2022 for first semester recovery interventions), as well as from two terms prior (e.g., spring 2021).[14] See Appendix C for more detail on alternative model specifications -- including the use of different functional forms and measures of treatment participation -- and the placebo tests we conducted to check for signs of selection bias influencing our estimates.

4.2 Interview Analysis

Each interview was conducted by a team of two researchers and were audio recorded. After each interview, the researchers completed an interview summary form that captured what they had learned about the intervention in each section of the interview protocol (e.g., reflections about participation, dosage, outcomes, challenges faced, and plans for next year). The team then wrote case memos about each intervention, documenting emerging findings from the summary forms and including quotes from cleaned interview transcripts to establish a chain of evidence to support our claims. These memos focused primarily on how the participants' account of intervention participation, dosage, content, and delivery might explain the results in the quantitative data. Upon completing the memos, the research team reviewed them to identify common themes across districts and interventions.

[14] For some districts, we are also able to include state standardized test scores in the value-added model, enabling us to include students with missing MAP Growth scores if they have non-missing state test scores from the same term. See Appendix for more detail.

These supplemental interviews elaborate on our quantitative findings, but they also have important limitations. Most notably, we interviewed leaders in only three districts that managed to start providing interventions to students during the 2021-22 school year and to collect data on students' participation. So, we cannot capture the range of implementation conditions faced by the districts that could not start interventions or collect data on them in the 2021–22 school year. Even in the districts where we conducted interviews, we did not capture the perspective of front-line implementers (e.g., teachers, tutors, interventionists). Instead, we rely on the perspective of central office leaders. In the end, the weight of the study and its conclusions rests on the quantitative impact analysis, while our qualitative findings help suggest the complexity surrounding the implementation of academic recovery interventions during the pandemic.

5. Results

5.1 Intervention Impacts

Table 4 and Figure 1 show the estimated impacts of treatment on math achievement for each of a series of math interventions in the four districts. We report impact estimates for each of the math interventions used across the four districts as the total effect across all grades served by the intervention and separated into effects for the elementary and middle school grade ranges served by the intervention when possible. In column 1, we report the coefficient on the indicator of whether a student received at least one session of treatment with math achievement as the outcome. For five of the resulting seven district/intervention combinations (across grades), the confidence interval for the impact includes zero, implying that we could not reject the null hypothesis of no impact. The confidence interval for all but one of these combinations also rules out effects larger than 0.05 standard deviations, a threshold under which school year intervention effect sizes are considered "small" in education research (Kraft, 2020). In the remaining two

cases, we estimate marginally significant impacts of participation on math achievement for all grades or a subset of grades: District A Tutoring/Small Group #1 and District B Virtual Learning. Though statistically significant, the magnitude of these estimated effects are also small, ranging from 0.02 to 0.04 standard deviations.

Column 2 shows coefficients from corresponding placebo tests, which examine selection bias by estimating the impact of participating in a subject-specific intervention (which plausibly only affects test achievement in that subject) on achievement in the opposite subject. In only one of the two cases in which we found small positive coefficients on participation in math intervention(s) did the intervention also pass the placebo test: District A Tutoring/Small Group #1. While it is possible the positive placebo estimate for District B Virtual Learning is representative of true impacts of the intervention on reading achievement, we also cannot rule out the possibility that students who participated in this intervention were different from students who did not participate in unobservable ways that led to their gains in both math and reading (as opposed to the fact that they participated in the intervention). Therefore, the positive placebo test reduces our confidence that the significant impact estimates for District B Virtual Learning should be directly attributed to the intervention.

Columns 3 and 4 show the estimated treatment effect per hour of treatment, along with its corresponding placebo test. We calculate these estimates by dividing the results in columns 1 and 2 by the average number of intervention hours for treated students in the district, reported in column 5. This approach, which assumes a linear relationship between treatment dosage and impact, is a fairly simplistic method of modeling the effect of an hour of treatment. We report these hourly estimates simply to convert impact estimates to a scale that is comparable across

interventions, given the considerable variation in the average treatment dosage received across interventions and districts.[15]

For context, we also report in column 6 the estimated impact we would have expected to see if the interventions had the same impact per hour as found in the prepandemic research on high-quality tutoring (Nickow et al., 2024; see Appendix C for additional detail). These "expected" total impacts for participating students range from 0.02 to 0.10 standard deviations across interventions. In all but one case (District B Virtual Learning grades 6-8), the expected impacts exceed the observed treatment effects. Furthermore, Figure 1 shows that, in most cases, the upper bounds of the confidence intervals for the treatment effects are below the expected effect estimate. In other words, we can rule out that the interventions had the same effect on math achievement per hour as the high-quality prepandemic tutoring programs in Nickow et al.'s (2024) meta-analysis.

Table 5 and Figure 2 show comparable results for seven district/intervention combinations (across grades) targeted at reading achievement. In only one case (District C Tutoring/Small Group #1), the estimate for the effect of any participation was statistically different from zero—but the point estimate was negative.[16] For the estimated effects of an hour of treatment, District A Tutoring/Small Group #1, District A Tutoring/Small Group #2, and District C Tutoring/Small Group #1 had significant impacts, though District C's intervention's impact was again negative.

[15] We estimate hourly effects in this way out of concern that our continuous measures of intervention participation are endogenous to student motivation or ability (discussed more in Appendix C). By dividing by the average dosage received across *all* students, we are able to calculate impact estimates that are comparable across interventions with less potential for bias.

[16] We think the negative point estimate on District C Tutoring/Small Group #1 (K-3) is likely to be caused by selection bias—according to conversations with administrators in that district, teachers frequently assigned students to participate in the program throughout the semester if that student was having difficulties with reading.

Because of the small, negative, and/or null effects estimated for each intervention, we did not estimate interaction effects of interventions for students who participated in multiple interventions within the year. Nevertheless, a small proportion of students received multiple ELA interventions in two of the four districts and math interventions in three of the four districts. The percentage of students receiving multiple interventions in a subject in these districts ranged from 5 to 22 percent. A higher percentage of students were receiving at least one intervention in both math and ELA, ranging from 14 to 74 percent across the four districts.

When we consider the specifics of participation in these interventions, the estimated impacts shown in Figures 1 and 2 are unsurprising. The number of students served and the amount of instruction provided were nearly always lower than planned (see Appendix Tables A3 and A4 for eligibility, participation, and dosage rates for math and ELA interventions). For example, districts' tutoring and small group interventions intended to serve between 5 and 45 percent of students across targeted schools and grades. However, over the course of the school year, the data indicate that these programs generally reached less than 20 to 30 percent (and sometimes less than 10%) of their intended enrollment, totaling 5 to 10 percent of all students in the targeted schools and grades.

The dose of programming students received also fell short of districts' plans. We found districts that had planned on offering students between 15 and 30 hours of mathematics tutoring per term (30 to 60 hours per year) ended up, on average, providing students 5 to 10 hours of math tutoring. For students who did participate, the number of sessions and the length of sessions were also often less than originally planned. In one district that had planned to offer students 90 sessions of tutoring over the course of the school year, students attended 13 sessions on average. In another district, math tutoring sessions were supposed to provide 100 minutes of instruction

during the week over five sessions; in practice, the average student attended 28 minutes of tutoring per week.

5.2 Intervention Implementation

The lack of impact from the interventions is unsurprising given the major implementation challenges identified in interviews with district leaders. Leaders mentioned a range of implementation challenges. All three districts reported challenges related to (a) reaching the targeted students consistently and equitably across schools; (b) staffing and staff capacity; (c) scheduling and delivering intervention services; (d) adapting interventions to accommodate existing federal, state, and district policies; and (e) building central office capacity and internal systems for scaling interventions. Importantly, each of these challenges was situated in and often exacerbated by the challenging context of the ongoing pandemic during the 2021-2022 school year.

Reaching Target Students

The interventions we studied typically targeted students based on one or more test performance thresholds (e.g., students who had scored below the 20th percentile on the MAP Growth test). Some interventions incorporated other eligibility criteria, such as low attendance rates, low course grades, or teacher recommendations when assigning students to interventions. But intervention leaders said they often decentralized decisions about student participation to schools and classrooms–effectively letting school personnel refer students to treatment–in the hope that the approach would generate buy-in from principals and teachers and help match students with appropriate interventions. In practice, this left principals and teachers to decide the balance between district-mandated eligibility criteria and their own professional judgment about which students had the greatest needs and/or would benefit most from the intervention.

Decentralizing eligibility decisions played out in several ways. For example, leaders of one intervention reported that teachers recommended students with test scores above the eligibility threshold because the teachers believed their students' scores were inflated and did not accurately reflect their achievement. While these teachers may have had a better understanding of their students' needs than what was reflected in test results, in other places, leaders reported that local decision-makers were directing services away from target populations and towards students with lower academic needs. Leaders of a reading intervention in one district reported schools focused on "bubble" students on the cusp of proficiency, rather than the low-performing students the intervention intended to serve (the intervention targeted students who performed at or below the 15th percentile of the school's test score distribution). In another district, 31% of the students who took part in a math intervention intended for students at or below the 20th percentile in math had scores *above* the 40th percentile. In two of the districts, leaders reported that schools occasionally used tutoring to help students who were performing at grade level but struggling with a specific topic. One leader concluded, "I think it [tutoring] is happening with the wrong set of kids."

Sometimes schools did not adhere to the intervention's targeting criteria because teachers believed the intervention was misaligned with student needs. For example, a leader of a math intervention in one district explained that some schools found that the students initially chosen for the intervention did not have the foundational skills necessary to benefit from it. In response, the district expanded its eligibility for the intervention from the lowest 25 percent of math performers to the lowest 30–35 percent of performers and gave teachers discretion to identify the students in this group who they thought would benefit from the intervention.

In another case, district leaders required schools to use district-level eligibility criteria (e.g., test score thresholds) for an initial wave of students and then allowed schools to use their own criteria to identify a second wave of students to access the intervention and fill in any available slots. Here, the district leaders felt this approach improved local buy-in and allowed schools to expand access to the intervention for more students while still preserving the district's interest in serving priority students. In the end, guidelines for assigning students to interventions that appear routine on paper were, in practice, hard to apply consistently.

Hiring and Deploying Staff

Districts used a range of strategies to staff interventions. Some contracted with vendors or hired new intervention specialists to work in schools. Others hired graduate assistants, retired and current teachers, or undergraduate and high school students. When possible, districts leveraged existing staff and existing relationships with vendors, individual volunteers, and community-based organizations to find intervention staff. Each approach presented its own challenges.

For example, districts that contracted with vendors gave up some control over the staff selection process, making it difficult for district leaders to ensure staff quality and consistency throughout the year. In a tutoring program that relied on community providers, the intervention leader said they felt like they did not have the luxury to do anything beyond basic background checks because of a tight labor market. Conversely, when districts hired intervention specialists and tutors directly, central offices—already stretched thin—had to invest substantial time and resources in the hiring process.

District leaders reported that leveraging existing staff and prior vendor relationships helped get interventions started earlier in the year. Starting from scratch, however, created delays in some cases. For example, leaders in one district said they spent the first five months of the

school year negotiating contracts with tutoring vendors to ensure that they were federally compliant and could be paid using ESSER funding. This meant that the district's tutoring programs did not launch until February and March 2022. In another district, a small team in the central office was responsible for hiring, onboarding, and training tutor providers. The leader of this team said its limited capacity created a bottleneck that delayed tutors' placement in schools. Once in schools, tutors had to work with teachers to identify student needs, delaying the delivery of services even further. In certain schools, persistent teacher turnover caused still other delays, as teacher–tutor relationships had to be restarted with each new hire.

Even when districts were able to get providers in place, other staffing problems could occur. One intervention leader reported needing to redeploy intervention specialists to cover regular classrooms because of COVID-related teacher absences during the Delta and Omicron surges. The leader of a reading intervention in another district concurred, explaining how the Delta surge affected staffing in one school:

> *At the start of the year, at one of our schools, they had something like 24 teachers out. They all had COVID. That was two weeks where interventionists were pulled from what they would regularly do. There's no way around it...you need a body in the classroom.*

"Usually, it was a domino effect," the leader said, with illnesses delaying interventions for weeks. In the same district, teachers reportedly used interventionists at the beginning of the year to help get small groups going, rather than delivering academic interventions. As one leader put it, the interventionists "have an eye on what the school needs," beyond their specific responsibilities to individual students.

Just as schools sometimes struggled to provide interventions because of teacher absences during the Delta and Omicron surges, COVID outbreaks also resulted in student absences that

could reduce the planned-for frequency and dosage. As students moved in and out of school and experienced stress and pressure related to the pandemic, some interventionists reported challenges with student behavior that made it harder to deliver the planned dose of academic support. Commenting on the amount of time spent in intervention sessions to manage student behavior, one district leader said, "If behavior is the thing that students need to get going [in school], maybe behavior should be the intervention." Finally, interviewees noted that even the fear of COVID could affect implementation. Early in the school year, for example, leaders said that some teachers were reluctant to send students to pull-out groups because they thought it would increase everyone's risk of infection.

Scheduling and Delivering Interventions

Interviews suggested that scheduling challenges could also make it harder for schools to deliver interventions as planned. "It comes down to access," said one intervention leader. "How easy is it to pull a student [from class] and bring them back?" Across all three districts, intervention leaders reported that delivering pull-out programs during the school day could be challenging. This was due, in part, to instructional time being fully planned out during the regular school day. Responding to data showing low intervention uptake and dosage, one district leader shared, "All of our literacy minutes were already being used for other things, so the data do not shock me." According to intervention leaders, some classroom teachers resisted pull-out interventions because they did not want students to miss grade-level core instruction. In other cases, students who would have been eligible for a pull-out intervention based on their test scores could not receive it because it conflicted with other, higher priority (or state-mandated) supports (e.g., ELL/Individualized Education Program services).

In multiple cases, leaders reported that intervention providers had to navigate schedules with individual teachers to meet with target students. This process meant that the same

intervention could occur at different times in different buildings, so the untreated counterfactual (what students missed during their intervention) varied across students and schools. One tutoring program director likened scheduling to a complex puzzle, a "game of figuring out where each person goes and fits [so that]…kids get hours but also we want tutors to get their hours." Local complexity and discretion sometimes meant that "schools did their own thing [when it came to scheduling] and that is hard for us [the district] to control," according to one district leader.

District-level schedules could also make accessing interventions easier or harder. For example, one district mandated extra intervention minutes for reading in all elementary school schedules, but not for math. As a result, reading intervention providers (a position that predated the pandemic) were reportedly more likely to find time to work with students than math intervention providers (a new position).[17]

In each of these cases, ease of scheduling was a function of who was responsible for scheduling and the extent to which intervention times aligned with existing school schedules. When intervention time was accounted for in school schedules and building administrators helped prioritize and coordinate scheduling, intervention leaders reported fewer scheduling issues. When schools worked directly with external contractors to schedule interventions outside of school hours, district leaders reported fewer issues and constraints. However, scheduling intervention sessions after the school day limited access for students who wanted to participate in extracurricular activities or did not have access to transportation after school.

[17] District leaders also observed programs having more success when they predated the pandemic because they could draw on existing relationships between interventionists and teachers and were well aligned with the school's core curricula. One district leader commented that schools with tutoring programs before the pandemic were much better positioned to grow their programs during the pandemic than those without prior programming.

Arranging intervention times was not the only scheduling challenge. In some cases, schools did not have adequate space for interventionists to work with students in small groups, further complicating intervention delivery. A district leader of a math intervention, for example, said:

> *Location was often an issue. Classrooms were not physically designed to have a group pulled in the back in many schools. So, their [students'] time was less because they lost minutes coming and going to the group.*

By contrast, in cases where intervention providers had space to work and could easily bring all their materials into the classroom, schools were reportedly better able to provide the planned dose of the intervention.

Aligning with Existing Federal, State, and District Policies

District leaders also faced the challenging task of embedding interventions in an existing system of federal, state, and local policies. At times, this required adapting interventions to accommodate existing rules and procedures, which, in turn, delayed the rollout of services or diminished their quality. To use ESSER funding for a tutoring intervention, leaders in one district had to revise their vendor contracts to meet federal contracting requirements, which delayed the intervention's rollout. Another district leader discussed having to comply with a state mandate requiring the use of tutoring to deliver a remediation curriculum, even though the leader believed it was more appropriate to use tutoring for grade-level content. Districts were implementing concurrent interventions that could conflict with the academic recovery intervention in ways confusing to teachers. To prevent confusion and frustration, district leaders prioritized aligning the features of the interventions and occasionally had to depart from evidence-based practices. For example, one district administrator discussed increasing tutoring

group sizes to more than what is considered best practice to align with the small-group sizes prescribed by the district's recently adopted, multi-tiered system of supports (MTSS) program.

Competing district initiatives also strained educator capacity for implementing interventions. Examples of concurrent initiatives implemented by the districts in the 2020-21 school year included new core curricula in reading and math, new training for teachers, COVID quarantine and testing policies and procedures, other digital tools for assessing and remediating student learning, new social-emotional and mental health supports, and other districtwide interventions. One district leader asked rhetorically, "How much capacity do people have? It [the multiple initiatives] is so much," implying that educators were overburdened and exhausted by the new policies and interventions adopted by the district. Another district leader said that, because schools were still learning how to implement other interventions that served the same student population as their tutoring program, it made it harder to ensure consistent scheduling for students and tutors.

Ensuring Central Office Capacity to Support Scale

Finally, district central offices often lacked capacity to oversee and coordinate the implementation of the interventions. Many of the representatives we spoke to worked in small teams, consisting of two or three total staff members, who were suddenly in charge of hiring intervention providers, coordinating school schedules, and overseeing implementation of an intervention for their entire district. Therefore, district leaders had limited time and capacity to manage these processes while also fulfilling other professional roles and responsibilities in the district. In reflecting on the past year, one district administrator shared that they could have provided better professional development to interventionists had it not been for the hours of new literacy training required by state law that they also had to provide for teachers.

District representatives also described working with internal systems that were not designed to handle the demands of interventions on such a large scale. As noted earlier, one district's process for hiring, onboarding, and training tutors was time consuming and delayed student placement with tutors. Another district leader shared that compliance management of diverse tutoring providers was cumbersome, primarily because the district did not have internal data systems to track tutoring hours and attendance across different providers. These remarks suggest that, to implement interventions at scale, districts need the authority and resources to invest in central office staffing and internal systems for overseeing these programs.

In summary, COVID-recovery interventions were often not implemented at the frequency or dosage originally planned in part because schools faced challenges related to reaching the targeted students, staffing, scheduling interventions, and limited central office capacity. Of course, these tasks were challenging because schools were attempting to help students recover from COVID while the pandemic was still happening. In addition, district leaders had limited capacity and systems from within the central office to take these interventions to scale, and sometimes had to adapt interventions to accommodate existing policies in ways that delayed services or reduced the quality of services offered to students.

The findings from our interviews underscore the challenging reality of the districts' implementation contexts. While our findings illuminate how these challenges hindered implementation in the 2021-2022 school year, many of the districts have already developed plans to address some of these persistent challenges in the 2022-2023 school year. Our interviews with district leaders suggest that implementation of recovery interventions is an iterative process that will require continual adjustments to internal (e.g., staffing shortages, intervention eligibility criteria and assignment policies, school schedules) and external (e.g., a surge in COVID-19

cases, state and federal policies) factors

6. Discussion & Conclusion

Consistent with other recent evidence that districts made little progress toward academic recovery on average during the 2021–22 school year (e.g., Jacobson, 2022; Kuhfeld & Lewis, 2022), our analysis of four districts' recovery interventions finds they served few students and had minimal (if any) positive effects on student achievement relative to business as usual. Of course, in theory, the wide range of catch-up efforts in these districts could be raising achievement for *all* students, making it hard to detect treatment effects from the interventions. These districts vary in the amount of recovery they need to return to their prepandemic achievement levels, but they all have more ground to make up than the average U.S. district, whose 2022 state test scores declined -0.49 grade levels in math and -0.31 grade levels in reading relative to 2019 (Reardon et al., 2023). As displayed in Table 6, the 2022 scores of the four districts herein declined by -0.22 to -0.56 grade levels in math and -0.08 to -0.60 grade levels in reading (Reardon et al., 2023). To catch up, student learning will need to move at a faster pace than it did prepandemic.

To better understand our findings on intervention participation, dosage, and impacts, we interviewed a subset of district leaders about implementing interventions. The results suggest that staffing and scheduling problems often plagued recovery efforts. As a result, many interventions served fewer students than originally intended—and often served students who were not in the targeted groups. In some cases, academic interventions displaced regular classroom instruction, reducing the contrast between the intervention "treatment" and business as usual, again making it difficult to detect treatment effects. Schools and districts alike experienced

the benefits and limits of decentralized decision-making, which can support local adaptation but also create inconsistency and confusion.

The implementation challenges district leaders recounted suggest that the simple-sounding logic of academic intervention—identify students in need and provide them extra support—belies a host of complex design decisions and implementation dynamics. Under existing decentralized decision-making structures and constraints on capacity and time, there are no easy solutions to address pandemic losses.

Providing sufficient intervention for all students in need is going to require historic action. States and districts can help by providing transparent and accessible measures of students' academic progress and recovery to schools, families, and students. Recent surveys indicate that parents currently underestimate the extent to which their own students are behind (Anderson et al., 2022; Hubbard & Burns, 2022). Districts and states may need to do more to inform families and communities about how students are doing now, whether they are on track for recovery, and what can be done if recovery does not look like it is happening at an adequate pace. There is evidence, outside of the pandemic context, that better alignment between grades and measured test scores results in better student achievement (Gershenson et al., 2022). In light of emerging evidence of grade inflation during the pandemic (Goldhaber & Young, 2023), it is important for school districts to make sure grades and other student outcomes (e.g., math and reading assessments) are aligned, given that grades are arguably the most direct means for schools to communicate with parents about student learning. Many schools are also implementing voluntary interventions which require school systems to articulate the extent to which students need supplemental (outside of the regular school day) services and to nudge families to use the intervention(s) to get even moderate student take-up (Robinson et al., 2022).

Successfully increasing the scale of interventions in districts will, in some cases, require more resources (e.g., staff and staff compensation). We show elsewhere that learning losses varied across districts (Goldhaber et al., 2022a) and that the ESSER funds districts received may be sufficient for recovery in low-income districts that were in person during the 2020-21 school year. But ESSER funds are unlikely to be sufficient for the larger share of districts that spent more time in remote status. Moreover, because the ESSER dollars were based on district poverty rates, these federal dollars will also be inadequate in the low-poverty districts that were remote for much of 2020-21 (Goldhaber et al., 2022b; Shores & Steinberg, 2022). In addition to funding, our findings suggest that districts may need to invest in central office capacity and internal administrative systems (e.g., data systems, hiring procedures) to implement academic recovery interventions at scale.

Given that a tight labor market limited the ability of schools to implement some recovery initiatives, districts may also need to cast a broader net to recruit adults to provide interventions in schools and seek out new or expanded partnerships with external organizations. Our interviews indicate that some districts managed to supplement their academic interventions with external partnerships. They tapped local community centers, educator preparation programs, college students, parents, and local community members to provide academic help. Given the scale of the need, these types of external partnerships are a key resource for expanding recovery efforts in the 2022-23 school year. Not every district we studied, however, was able to leverage external partnerships to support academic recovery.

Finally, districts will need help to expand their interventions to be commensurate with their students' losses. In most cases, this will mean expanding student participation and dosage in existing programs, as well as *layering* interventions (e.g., high dosage tutoring and an extended

school year) for targeted students. To illustrate this point, we end the paper by translating the average student's remaining recovery in these districts to the hours of high dosage tutoring that would be needed for a full recovery (see Table 6). Just for students in grades 3-8, the four districts in this study will need to deliver an average of 9 to 23 hours of math tutoring per student in addition to an average of 2 to 21 hours of reading tutoring per student to fully recover all students. In these large districts, this roughly equates to between 150,000 and 650,000 total hours of reading tutoring and between 370,000 and 1,430,000 total hours of math tutoring provided by a district to students in grades 3-8. If we assume tutors work 5-hour days for 180 days a year, delivering 150,000 hours of reading tutoring would require deploying around 160 reading tutors. For most districts, this level of intervention would be a significant step up in intensity from what was implemented during the 2021–22 school year. Districts do not, however, have to tackle this problem alone. States and other civic leaders can help districts mobilize communities by providing information, political cover (for example, on extending learning time), and investing in the capacity of districts, schools, and communities to support and advocate for recovery. A coordinated approach is not only important for school systems' response to pandemic-related learning disruptions, but will also inform our responses to future emergencies that disrupt schooling for extended periods of time.

Complete academic recovery—and, ideally, academic acceleration—is as urgent as it is challenging. Especially in the places hit hardest by the pandemic, academic recovery from COVID-19 is likely to require an all-hands-on-deck response for the next several years. Recovery is unlikely to be completed when the federal dollars run out in September 2024, suggesting that states will need to take further action to support additional academic interventions.

References

Anderson, M., Faverio, M., & McClain, C. (2022, June 2). *How teens navigate school during COVID-19*. Pew Research Center. https://www.pewresearch.org/internet/2022/06/02/how-teens-navigate-school-during-covid-19/#teens-and-parents-express-their-views-about-virtual-learning-and-the-pandemic-s-impact-on-educational-achievement

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher, 48*(4), 217-228.

Barry, S. S., & Sass, T. R. (2022). The impact of a 2021 summer school program on student achievement. Georgia Policy Labs Report. https://gpl.gsu.edu/publications/impact-of-a-2021-summer-school-program-on-student-achievement/

Bradshaw, C. P., Kush, J. M., Braun, S. S., & Kohler, E. A. (2023). The perceived effects of the onset of the COVID-19 pandemic: A focus on educators' perceptions of the negative effects on educator stress and student well-being. School Psychology Review, 1–14. https://doi.org/10.1080/2372966X.2022.2158367

Camera, L. (2022, January 18). Country's biggest school districts resist going remote as closures spread nationwide. *U.S. News & World Report*. https://www.usnews.com/news/education-news/articles/2022-01-18/countrys-biggest-school-districts-resist-going-remote-as-closures-spread-nationwide.

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

Cortes, K., Kortecamp, K., Loeb, S., & Robinson, C. (2023). A scalable approach to high-impact tutoring for young readers: Results of a randomized controlled trial. National Student Support Accelerator Report. https://studentsupportaccelerator.org/sites/default/files/Scalable%20Approach%20to%20High-Impact%20Tutoring.pdf

Curriculum Associates (2020). *Understanding student needs: Early results from fall assessments*. [Research Brief]. Curriculum Associates.

DeArmond, M., Hill, P., Destler, K., & Campbell, C. *Whack-a-mole: School systems respond to disrupted learning in 2021*. Center on Reinventing Public Education. https://crpe.org/whack-a-mole-school-systems-respond-to-disrupted-learning-in-2021/

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

Domash, A., & Summers, L. H. (2022). *How tight are U.S. labor markets?* (Working Paper No. 29739). National Bureau of Economic Research. https://www.nber.org/papers/w29739

Dorn, E., Hancock, B., & Sarakatsannis, J. (2021, July 27). *COVID-19 and education: The lingering effects of unfinished learning*. McKinsey & Company. https://www.mckinsey.com/industries/education/our-insights/covid-19-and-education-the-lingering-effects-of-unfinished-learning

Doty, E., Kane, T. J., Patterson, T., & Staiger, D. O. (2022). *What do changes in state test scores imply for later life outcomes?* (No. w30701). National Bureau of Economic Research.

Education Policy Innovation Collaborative (EPIC). (2021). *K–8 Student achievement and achievement gaps on Michigan's 2020–21 benchmark and summative assessments*. https://epicedpolicy.org/wp-content/uploads/2022/01/EPIC_BenchmarkII_Rptv1_Dec2021.pdf

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.

Gershenson, S., Holt, S., & Tyner, A. (2022). *Making the Grade: The Effect of Teacher Grading Standards on Student Outcomes.* (IZA Discussion Paper No. 15556). https://ssrn.com/abstract=4226363

Goldhaber, D., Kane, T. J., McEachin, A., & Morton, E. (2022a). *A comprehensive picture of achievement across the COVID–19 pandemic years: Examining variation in test levels and growth across districts, schools, grades, and students*. (CALDER Working Paper No. 266–0522). https://caldercenter.org/sites/default/files/CALDER%20Working%20Paper%20266-0522_0.pdf

Goldhaber, D., Kane, T.J., McEachin, A., Morton, E., Patterson, T., & Staiger, D.O. (2022b). *The Consequences of Remote and Hybrid Instruction During the Pandemic.* (Working Paper No. 30010). National Bureau of Economic Research. https://www.nber.org/papers/w30010

Goldhaber, D. & Young, M.G. (2023). Course grades as a signal of student achievement: Evidence on grade inflation from before and after COVID-19. CALDER Policy Brief No. 35.https://caldercenter.org/publications/course-grades-signal-student-achievement-evidence-grade-inflation-and-after-covid-19

Hamilton, L., & Gross, B. (2021). How has the pandemic affected students' social-emotional well-being? A review of the evidence to date. *Center on Reinventing Public Education*.

Hanushek, E. A. (2023, October 6). *Generation lost: The pandemic's lifetime tax.* Education Next. https://www.educationnext.org/generation-lost-the-pandemics-lifetime-tax/

Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes:

Combining effects with costs. Educational Evaluation and Policy Analysis, 31(1), 3–29.

Hubbard, B., & Burns, A. (2022, June). *Hidden in plain sight: A way forward for equity-centered family engagement*. Learning Heroes. https://learningheroes.wpenginepowered.com/wp-content/uploads/2022/06/Parents22-Research-Deck-1.pdf

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

Jacobson, L. (2022, October). Exclusive literacy data: Small gains since last fall, but no reading rebound. *The74*. https://www.the74million.org/article/exclusive-literacy-data-small-gains-since-last-fall-but-no-reading-rebound/

Jones, S. E., Ethier, K. A., Hertz, M., DeGue, S., Le, V. D., Thornton, J., Lim, C., Dittus, P., & Gede, S. (2022). Mental health, suicidality, and connectedness among high school students during the COVID-19 pandemic (MMWR Suppl 2022;71(Suppl-3):16–21; Adolescent Behaviors and Experiences Survey, United States, January-June 2021). http://dx.doi.org/10.15585/mmwr.su7103a3external

Jordan, P. W., & DiMarco, B. (2022). *Educators and ESSER: How pandemic spending is reshaping the teaching profession*. FutureEd at Georgetown University's McCourt School of Public Policy. https://www.future-ed.org/educators-and-esser-how-pandemic-spending-in-reshaping-the-teaching-profession/

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* (NBER Working Paper No. 14607). National Bureau of Economic Research.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* [Research Paper]. MET Project. Bill & Melinda Gates Foundation.

Kingsbury, G. G., Nesterak, M., & Freeman, E. (2014, March). The potential of adaptive assessment. *Education Leadership, 71*(6). https://www.nwea.org/research/publication/the-potential-of-adaptive-assessment/

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

Kuhfeld, M., & Lewis, K. (2022). *Student achievement in 2021–22: Cause for hope and continued urgency*. NWEA. https://www.nwea.org/research/publication/student-achievement-in-2021-22-cause-for-hope-and-continued-urgency

Kuhfeld, M., & Soland, J. (2021). The learning curve: Revisiting the assumption of linear growth during the school year. *Journal of Research on Educational Effectiveness, 14*(1), 143-171.

Kuhfeld, M., Soland, J., Lewis, K., & Morton, E. (2022, March 3). The pandemic has had devastating impacts on learning. What will it take to help students catch up? *Brookings*.

https://www.brookings.edu/blog/brown-center-chalkboard/2022/03/03/the-pandemic-has-had-devastating-impacts-on-learning-what-will-it-take-to-help-students-catch-up/

Lewis, K & Kuhfeld, M. (2023). *Education's long COVID: 2022-2023 achievement data reveal stalled progress toward pandemic recovery*. NWEA. https://www.nwea.org/research/publication/educations-long-covid-2022-23-achievement-data-reveal-stalled-progress-toward-pandemic-recovery/

Lewis, K., Kuhfeld, M., Ruzek, E., & McEachin, A. (2021). *Learning during COVID-19: Reading and math achievement in the 2020–21 school year*. NWEA. https://www.nwea.org/content/uploads/2021/07/Learning-during-COVID-19-Reading-and-math-achievement-in-the-2020-2021-school-year.research-brief-1.pdf

Lynch, K., An, L., & Mancenido, Z. (2022). *The impact of summer programs on student mathematics achievement: A meta-analysis*. (EdWorkingPaper No. 21-379). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai21-379

McEachin, A., Augustine, C. H., & McCombs, J. (2018). Effective summer programming: What educators and policymakers should know. *American Educator*, *42*(1), 10. https://eric.ed.gov/?id=EJ1173313

McEachin, A., Welsh, R., & Brewer, D. J. (2016). Student achievement within a portfolio management model: Early results from New Orleans. *Educational Evaluation and Policy Analysis, 38*(4), 669–691.

Meier, K. J., O'Toole, L. J., & Nicholson-Crotty, S. (2004). Multilevel Governance and Organizational Performance: Investigating the Political-Bureaucratic Labyrinth. *Journal of Policy Analysis and Management, 23*(1), 31–47.

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.

Opper, I. M., Park, R.J., Husted L. (2023). The effect of natural disasters on human capital in the United States. *Nature 7*(9), 1442-1453. doi: 10.1038/s41562-023-01610-z

Patrick, S. W., Henkhaus, L. E., Zickafoose, J. S., Lovell, K., Halvorson, A., Loch, S., Letterie, M., & Davis, M. M. (2020). Well-being of parents and children during the COVID-19 pandemic: A national survey. Pediatrics, 146(4), e2020016824. https://doi.org/10.1542/peds.2020-016824

Polikoff, M., & Houston, D. (2022, September). Experts Say Kids Are Far Behind After COVID; Parents Shrug. Why the Disconnect?. *The74m.* *https://www.the74million.org/article/experts-say-kids-are-far-behind-after-covid-parents-shrug-why-the-disconnect/*

Reardon, S. F., Fahle, E. M., Ho, A. D., Shear, B. R., Kalogrides, D., Saliba, J. & Kane, T. J. (2023). Stanford Education Data Archive (Version SEDA 2022 2.0). Retrieved from http://purl.stanford.edu/db586ns4974.

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. https://www.edworkingpapers.com/ai22-654

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.

Sandfort, J., & Moulton, S. (2015). *Effective implementation in practice: Integrating public policy and management*. Jossey-Bass

Schwartz, H. (2022, March 14). What is really polarizing schools right now? *Education Week*. https://www.edweek.org/leadership/opinion-what-is-really-polarizing-schools-right-now/2022/03

Shores, K., & Steinberg, M. P. (2022). Fiscal Federalism and K–12 Education Funding: Policy Lessons from Two Educational Crises. *Educational Researcher*, 0(0). https://doi.org/10.3102/0013189X221125764.

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. (2022a). *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. (2022b) *National Assessment of Educational Progress (NAEP) 2022 Mathematics and Reading Assessment*. Institute of Education Sciences, National Center for Education Statistics. https://www.nationsreportcard.gov/highlights/mathematics/2022/ and https://www.nationsreportcard.gov/highlights/reading/2022/

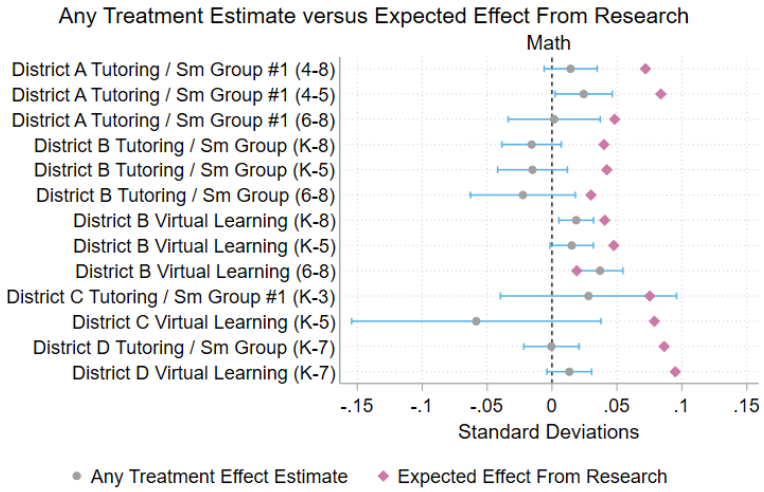
West, M. R., & Lake, R. (2021). *How Much Have Students Missed Academically Because of the Pandemic? A Review of the Evidence to Date.* Center on Reinventing Public Education.

White, S., Groom-Thomas, L., & Loeb, S. (2023). *A systematic review of research on tutoring implementation: Considerations when undertaking complex instructional supports for students.* (EdWorkingPaper 22–652). Annenberg Institute at Brown University

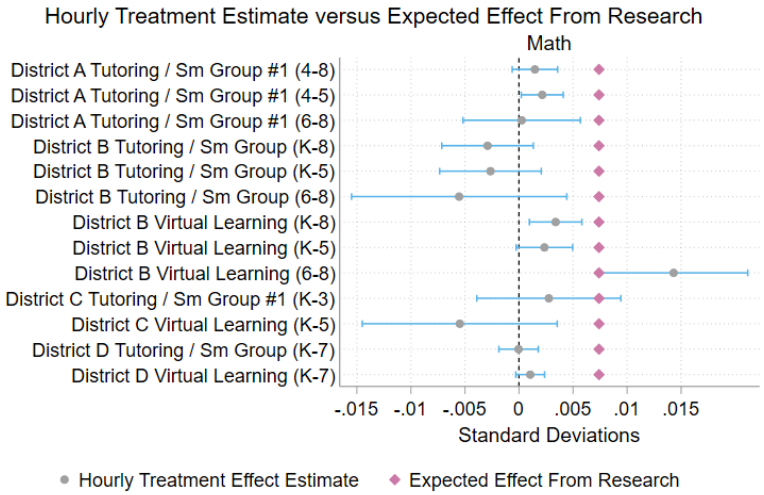
Figures and Tables

Figure 1. Estimated Treatment Effect of Math Interventions

A. Impact estimates for binary measure of treatment



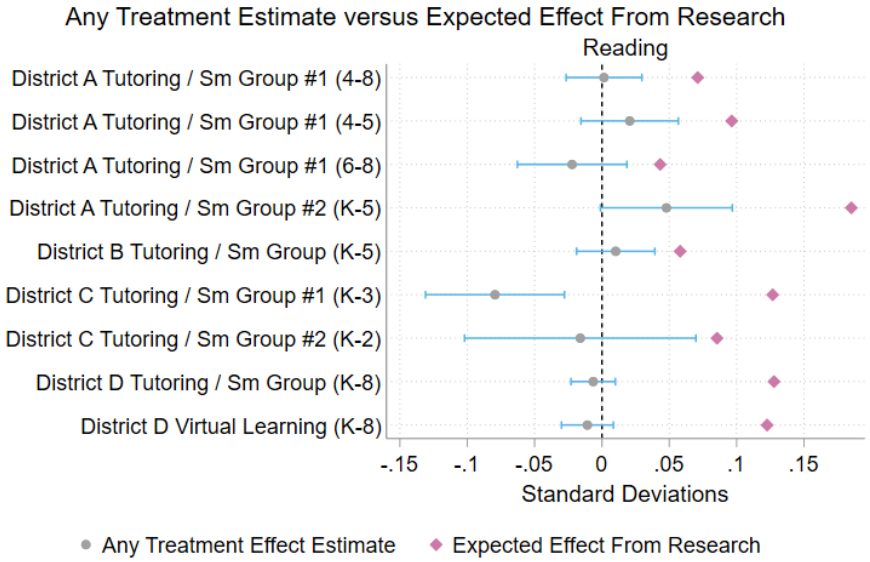
B. Impact estimates for hourly measure of treatment



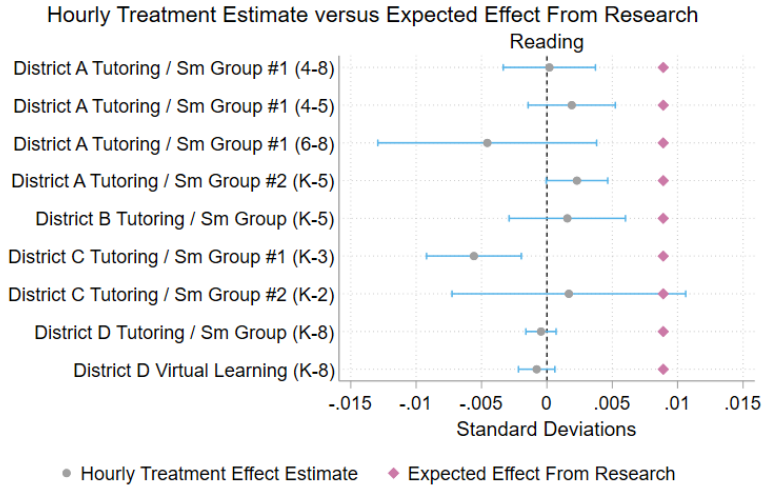
Notes: Point estimates (grey dots) show the average effect of receiving any amount of math intervention (panel A) or one hour of math intervention (panel B) in a given term on math MAP Growth scores at the end of that term. For all districts aside from District C the model used is a stacked model with a fall and spring term for each student; models for District C include a fall term only. Covariates in the model 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-term fixed effects. When applicable, models also include indicators for a student scoring below a certain MAP Growth threshold at baseline for interventions where eligibility is based on a MAP Growth score cutoff. Blue lines indicate 95% confidence intervals. The expected effect of treatment (pink diamonds) in panel A is calculated by multiplying the average dosage in hours by the estimated average hourly effect of high dosage tutoring in math (0.0074 SD) according to the meta-analysis by Nickow et al. (2024).

Figure 2. Estimated Treatment Effects of Reading Interventions

A. Impact estimates for binary measure of treatment



B. Impact estimates for hourly measure of treatment



Notes: Point estimates (grey dots) show the average effect of receiving any amount of reading intervention (panel A) or one hour of reading intervention (panel B) in a given term on reading MAP Growth scores at the end of that term. For all districts aside from District C the model used is a stacked model with a fall and spring term for each student; models for District C include a fall term only. Covariates in the model include participation indicators for other reading interventions and math 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-term fixed effects. When applicable, models also include indicators for a student scoring below a certain MAP Growth threshold at baseline for interventions where eligibility is based on a MAP Growth score cutoff. Blue lines indicate 95% confidence intervals. The expected effect of treatment (pink diamonds) in panel A is calculated by multiplying the average dosage in hours by the estimated average hourly effect of high dosage tutoring in literacy (0.0089 SD) according to the meta-analysis by Nickow et al. (2024).

Table 1. Sample Demographics

	Study Districts	Nationwide NWEA Districts	U.S. Public Schools
Average school enrollment	632	467	472
% FRPL	68%	54%	55%
% Asian	4%	4%	4%
% Hispanic	42%	21%	25%
% Black	25%	16%	15%
% White	26%	52%	49%
% City	77%	29%	28%
% Suburb	18%	28%	28%
% Town	0%	11%	12%
% Rural	5%	31%	32%

Note: FRPL=free or reduced priced lunch. The source of the variables is the Common Core of Data (CCD) collected by the National Center for Education Statistics during the 2019-2020 school year.

Table 2. Program Usage Across Sample Districts

	Tutoring and Small Group Interventions	Out-of-School Time	Virtual Learning	Extended Calendar
District A	X			X
District B	X	X	X	
District C	X	X	X	X
District D	X	X	X	

Table 3. Supplemental Implementation Interviews and Intervention Programs

	Number of Participants	Intervention Programs
District A	3	● Tutoring / small group intervention #1 (reading and math) ● Tutoring / small group intervention #2 (reading)
District B	3	● Tutoring / small group intervention (reading and math) ● Virtual learning program intervention (math)
District C	3	● Tutoring / small group intervention #1 (reading and math) ● Tutoring / small group intervention #2 (reading) ● Virtual learning program intervention (math)

Table 4. Estimated Treatment Effects of Math Interventions

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Any Participation		Hourly			
District	Intervention (Grades)	Sample students	% Treated	Point Estimate (SE)	Placebo Estimate (SE)	Estimated Impact (SE)	Placebo Estimate (SE)	Avg Dosage (Hours)	Expected Effect from Tutoring Research
A	Tutoring/Sm Group #1 (4-8)	43,270	6.07%	0.0143	0.0022	0.00147	0.00022	9.72	0.0719
				(0.0104)	(0.0139)	(0.00107)	(0.00143)		
	Tutoring/Sm Group #1 (4-5)	18,212	9.50%	0.0244*	0.0138	0.00215*	0.00122	11.32	0.0838
				(0.0112)	(0.0191)	(0.00099)	(0.00169)		
	Tutoring/Sm Group #1 (6-8)	25,058	3.54%	0.0016	-0.0054	0.00025	-0.00082	6.53	0.0483
				(0.0181)	(0.0194)	(0.00277)	(0.00298)		
B	Tutoring/Sm Group (K-8)	40,828	2.88%	-0.0157	-0.0034	-0.00291	-0.00063	5.40	0.0400
				(0.0117)	(0.0142)	(0.00216)	(0.00262)		
	Tutoring/Sm Group (K-5)	27,589	3.45%	-0.0151	-0.0048	-0.00264	-0.00085	5.70	0.0422
				(0.0137)	(0.0169)	(0.00240)	(0.00296)		
	Tutoring/Sm Group (6-8)	13,239	1.65%	-0.0225	-0.0066	-0.00554	-0.00164	4.06	0.0300
				(0.0206)	(0.0242)	(0.00508)	(0.00597)		
	Virtual Learning (K-8)	40,828	18.53%	0.0186**	0.0159	0.00339**	0.00290	5.49	0.0406
				(0.0068)	(0.0084)	(0.00124)	(0.00153)		
	Virtual Learning (K-5)	27,589	20.61%	0.0151	0.0032	0.00235	0.00050	6.42	0.0475
				(0.0085)	(0.0106)	(0.00133)	(0.00165)		
	Virtual Learning (6-8)	13,239	14.07%	0.0369**	0.037**	0.0143**	0.0143**	2.58	0.0191
				(0.0090)	(0.0123)	(0.00350)	(0.00477)		
C	Tutoring/Sm Group (K-3)	15,502	3.46%	0.0281	0.0332	0.00276	0.00326	10.17	0.0746
				(0.0346)	(0.0297)	(0.00340)	(0.00292)		
	Virtual Learning (K-5)	19,242	85.56%	-0.0584	-0.190***	-0.00548	-0.01784***	10.65	0.0828
				(0.0490)	(0.0520)	(0.00460)	(0.00488)		
D	Tutoring/Small Group (K-7)	20,926	5.58%	-0.0005	0.0089	-0.00004	0.00077	11.67	0.0864
				(0.0109)	(0.0136)	(0.00093)	(0.00116)		
	Virtual Learning (K-7)	20,926	24.60%	0.0133	0.02*	0.00104	0.00156*	12.83	0.0950
				(0.0088)	(0.0099)	(0.00068)	(0.00077)		

* p<0.05 ** p<0.01

Notes: Point estimates show the average effect of receiving any amount of math intervention in a given term on math MAP Growth scores at the end of that term, and the estimated effect of receiving one hour of math intervention. The estimated effect of receiving one hour is calculated by dividing the average effect of receiving any amount of intervention by the average number of hours received among treated students (column 5). For all districts

aside from District C the model used is a stacked model with a fall and spring term for each student; models for District C include a fall term only. Covariates in the model 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-term fixed effects. When applicable, models also include indicators for a student scoring below a certain MAP Growth threshold at baseline for interventions where eligibility is based on a MAP Growth score cutoff. Placebo estimates show the effect of any amount of math intervention on MAP Growth reading scores, using the same model specifications. Average dosage indicates the average number of hours treated students received the intervention for each term. The expected effect from tutoring research is calculated by multiplying the average dosage in hours by the estimated average hourly effect of high dosage tutoring in math (0.0074 SD) according to the meta-analysis by Nickow et al. (2024).

Table 5. Estimated Treatment Effects of Reading Interventions

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Any Participation		Hourly			
District	Intervention (Grades)	Sample students	% Treated	Point Estimate (SE)	Placebo Estimate (SE)	Estimated Impact (SE)	Placebo Estimate (SE)	Avg Dosage (Hours)	Expected Effect from Tutoring Research
A	Tutoring/Sm Group #1 (4-8)	37,333	6.18%	0.0015	0.0203*	0.00019	0.00254*	7.99	0.0711
				(0.0143)	(0.0093)	(0.00180)	(0.00117)		
	Tutoring/Sm Group #1 (4-5)	12,345	9.88%	0.0206	0.0168	0.00190	0.00155	10.82	0.0963
				(0.0184)	(0.0144)	(0.00170)	(0.00133)		
	Tutoring/Sm Group #1 (6-8)	24,988	4.38%	-0.0222	0.0240*	-0.00456	0.00494*	4.86	0.0433
				(0.0207)	(0.0119)	(0.00427)	(0.00245)		
	Tutoring/Sm Group #2 (K-5)	28,754	1.73%	0.0478	-0.0281	0.00230	-0.00135	20.80	0.1851
				(0.0250)	(0.0234)	(0.00120)	(0.00113)		
B	Tutoring/Sm Group (K-5)	17,964	5.40%	0.0102	0.0008	0.00156	0.00012	6.52	0.0580
				(0.0148)	(0.0136)	(0.00227)	(0.00208)		
C	Tutoring/Sm Group #1 (K-3)	15,533	3.45%	-0.0794**	-0.0236	-0.00558**	-0.00166	14.24	0.1203
				(0.0263)	(0.0230)	(0.00185)	(0.00162)		
	Tutoring/Sm Group #2 (K-2)	10,278	3.69%	-0.0161	-0.00081	-0.00168	-0.00088	9.60	0.0839
				(0.0438)	(0.0369)	(0.00456)	(0.00384)		
D	Tutoring/Sm Group (K-8)	22,686	4.70%	-0.0065	-0.0105	-0.00045	-0.00073	14.35	0.1277
				(0.0084)	(0.0078)	(0.00059)	(0.00054)		
	Virtual Learning (K-8)	22,686	13.80%	-0.0108	0.0128	-0.00078	0.00093	13.78	0.1226
				(0.0098)	(0.0089)	(0.00071)	(0.00065)		

* p<0.05 ** p<0.01

Notes: Point estimates show the average effect of receiving any amount of reading intervention in a given term on reading MAP Growth scores at the end of that term and the estimated effect of receiving one hour of reading intervention. The estimated effect of receiving one hour is calculated by dividing the average effect of receiving any amount of intervention by the average number of hours received among treated students (column 5). For all districts aside from District C the model used is a stacked model with a fall and spring term for each student; models for District C include a fall term only. Covariates in the model include participation indicators for other reading interventions and math 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-term fixed effects. When applicable, models also include indicators for a student scoring below a certain MAP Growth threshold at baseline for interventions where eligibility is based on a MAP Growth score cutoff. Placebo estimates show the effect of the any amount of reading intervention on MAP Growth math scores, using the same model specifications. Average dosage indicates the average number of hours treated students received the intervention for each term. The expected effect from tutoring research is calculated by multiplying the average dosage in hours by the estimated average hourly effect of high dosage tutoring in literacy (0.0089 SD) according to the meta-analysis by Nickow et al. (2024).

Table 6. Estimated Achievement Loss and Recovery from Spring 2019 to 2022, Grades 3-8

	Subject	Spring 2019 (SDs)	Spring 2022 (SDs)	Change from spring 2019 to spring 2022 (SDs)	Change from spring 2019 to spring 2022 (grade levels)	Avg high-dosage tutoring hours per student to eliminate the loss
District A	Math	-0.06	-0.21	-0.15	-0.49	21.6
	Reading	-0.31	-0.33	-0.02	-0.08	2.4
District B	Math	-0.09	-0.25	-0.16	-0.56	23.1
	Reading	-0.03	-0.20	-0.17	-0.60	20.5
District C	Math	0.00	-0.06	-0.06	-0.22	8.6
	Reading	0.05	0.02	-0.03	-0.10	3.6
District D	Math	0.19	0.04	-0.15	-0.51	21.6
	Reading	-0.02	-0.10	-0.08	-0.29	9.7

Notes. Spring 2019 and spring 2022 estimates are from the Stanford Education Data Archive (Version SEDA 2022 2.0; Reardon et al., 2023) 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. Tutoring hours to eliminate the loss are calculated based on Nickow et al.'s (2024) estimates that approximately 38.9 hours of tutoring per year in math results in a 0.27 SD gain in math achievement and 35.0 hours of tutoring per year in literacy results in a 0.29 SD gain in reading achievement.

Appendix A. Supplementary Tables

Appendix Table A1. NWEA MAP Math Scores by Term, Intervention, and Treatment Status

	Fall 2021 MAP RIT score				Spring 2022 MAP RIT score				Fall 2021 MAP Normed				Spring 2022 MAP Normed			
	Treated		Not Treated		Treated		Not Treated		Treated		Not Treated		Treated		Not Treated	
Intervention (Grades)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
District A Tutoring #1 (4-8)	192.47	14.97	209.19	18.33	201.50	15.75	217.48	18.56	-1.09	0.80	-0.26	1.01	-1.00	0.88	-0.23	1.02
District A Tutoring #1 (4-5)	187.47	12.73	201.09	16.76	198.26	15.27	212.36	17.70	-1.11	0.83	-0.22	1.09	-0.99	0.94	-0.14	1.08
District A Tutoring #1 (6-8)	202.65	14.01	214.84	17.23	207.93	14.67	221.11	18.30	-1.04	0.75	-0.29	0.96	-1.03	0.75	-0.29	0.97
District B Tutoring (K-8)	183.74	25.62	189.80	31.16	192.41	23.94	198.29	29.09	-0.74	0.91	-0.14	1.08	-0.88	0.96	-0.31	1.11
District B Tutoring (K-5)	177.37	22.81	176.60	26.69	187.13	22.17	187.08	25.61	-0.82	0.90	-0.14	1.10	-0.95	0.97	-0.35	1.13
District B Tutoring (6-8)	212.93	15.56	217.85	19.10	215.97	16.08	222.37	20.13	-0.38	0.84	-0.13	1.06	-0.56	0.84	-0.23	1.06
District B Assigned Software (K-8)	182.90	29.89	191.16	31.09	192.33	28.37	199.44	28.94	-0.44	1.22	-0.09	1.04	-0.59	1.26	-0.27	1.06
District B Assigned Software (K-5)	171.54	23.57	177.95	27.13	182.09	22.97	188.38	25.96	-0.60	1.24	-0.05	1.03	-0.77	1.28	-0.27	1.06
District B Assigned Software (6-8)	218.74	16.34	217.61	19.47	224.77	17.27	221.84	20.49	0.08	1.01	-0.17	1.06	0.00	0.99	-0.28	1.06
District C Tutoring (K-3)	162.71	13.69	176.72	17.21	173.43	13.24	187.77	16.48	-0.99	0.71	0.16	1.03	-1.02	0.77	0.07	1.01
District C Assigned Software (K-5)	184.45	21.01	170.45	17.23	194.36	19.75	177.20	14.83	0.23	1.00	-0.35	1.03	0.13	0.99	-0.50	0.99
District D Interventions (K-7)	182.05	27.05	194.70	29.77	193.61	24.14	206.50	27.06	-0.75	0.92	0.68	0.91	-0.68	0.96	0.60	0.90
District D Interventions (K-5)	172.60	24.53	185.54	26.70	186.40	22.86	198.74	24.90	0.76	0.92	0.71	0.95	-0.67	0.98	0.63	0.94
District D Interventions (6-7)	206.05	16.05	226.86	12.93	211.91	16.36	233.81	13.06	-0.70	0.93	0.57	0.74	-0.72	0.91	0.51	0.73
District D Classroom (K-7)	181.03	29.90	190.72	29.21	192.16	26.12	202.43	26.59	0.72	0.92	0.22	1.13	-0.68	0.95	0.18	1.10
District D Small Group (K-7)	194.42	15.95	189.44	30.24	204.02	16.30	201.27	27.41	-0.80	0.87	0.21	1.13	-0.73	0.92	0.18	1.10
District D Supervised (K-7)	181.07	26.35	193.91	29.86	192.77	23.57	205.62	27.12	-0.80	0.91	0.57	0.98	-0.73	0.95	0.50	0.96
District D Other (K-7)	186.82	26.61	190.00	29.54	197.26	23.20	201.73	26.88	-0.69	0.90	0.18	1.14	-0.66	0.94	0.15	1.11

Appendix Table A2. NWEA MAP Reading Scores by Term, Intervention, and Treatment Status

	Fall 2021 MAP RIT score				Spring 2022 MAP RIT score				Fall 2021 MAP Normed				Spring 2022 MAP Normed			
	Treated		Not Treated		Treated		Not Treated		Treated		Not Treated		Treated		Not Treated	
Intervention (Grades)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
District A Tutoring #1 (4-8)	189.56	15.96	204.55	18.98	195.18	16.19	209.13	18.59	-1.09	0.88	-0.32	1.10	-1.12	0.99	-0.37	1.13
District A Tutoring #1 (4-5)	185.01	14.31	197.96	17.87	193.52	15.67	205.47	17.80	-0.95	0.81	-0.16	1.06	-0.90	0.94	-0.15	1.09
District A Tutoring #1 (6-8)	194.54	16.18	207.55	18.71	196.97	16.55	210.81	18.71	-1.24	0.93	-0.39	1.11	-1.36	0.98	-0.46	1.13
District A Tutoring #2 (K-5)	175.91	12.02	179.85	24.48	188.56	14.24	190.39	23.23	-1.31	0.59	-0.23	1.09	-1.05	0.85	-0.25	1.11
District B Tutoring (K-5)	180.09	22.14	186.59	26.66	185.66	20.62	192.52	24.49	-0.54	0.87	-0.02	1.08	-0.75	0.96	-0.25	1.13
District C Tutoring #1 (K-3)	157.36	12.54	174.68	19.72	169.74	13.26	184.10	17.95	-0.88	0.67	0.20	1.03	-0.86	0.71	0.08	1.00
District C Tutoring #2 (K-2)	152.43	10.88	167.05	16.90	167.48	14.83	180.29	16.60	-0.81	0.72	0.21	1.06	-0.73	0.89	0.12	0.99
District D Interventions (K-8)	180.38	27.53	199.15	29.05	189.30	24.60	207.50	25.43	-0.63	1.02	0.69	0.91	-0.64	1.06	0.63	0.91
District D Interventions (K-5)	169.03	24.75	186.87	27.11	180.62	23.02	197.66	24.44	-0.58	0.98	0.74	0.98	-0.57	1.03	0.67	0.97
District D Interventions (6-8)	201.89	18.09	223.78	12.37	205.80	18.23	227.24	12.71	-0.74	1.07	0.59	0.72	-0.77	1.11	0.54	0.76
District D Classroom (K-8)	183.49	23.54	191.70	30.24	192.88	21.23	200.18	26.94	-0.62	0.98	0.18	1.15	-0.58	1.01	0.13	1.16
District D Small Group (K-8)	180.05	27.46	194.79	29.76	189.18	24.39	203.15	26.45	-0.62	0.99	0.37	1.10	-0.62	1.02	0.32	1.11
District D Supervised (K-8)	186.81	24.67	192.13	30.89	194.31	22.43	200.92	27.39	-0.72	1.02	0.32	1.10	-0.72	1.07	0.27	1.10
District D Other (K-8)	164.83	26.64	194.78	28.45	176.46	24.41	202.91	25.30	-0.74	0.99	0.24	1.13	-0.72	1.06	0.19	1.13

Appendix Table A3. Math Interventions Eligibility, Participation, and Dosage

Intervention (Grades)	Panel A: Eligible vs Treated			Panel B: Dosage	
	% eligible, not treated	% eligible, treated	% ineligible, treated	Intended dosage in hours (per term)	Average hours attended (per term)
District A Tutoring #1 (4-8)	37.00%	5.98%	0.10%	15	9.72
District A Tutoring #1 (4-5)	37.16%	9.31%	0.19%	15	11.32
District A Tutoring #1 (6-8)	36.88%	3.51%	0.03%	15	6.53
District B Tutoring (K-8)	24.46%	1.31%	1.56%	18.5-55.5	5.40
District B Tutoring (K-5)	25.16%	1.70%	1.75%	18.5-55.5	5.70
District B Tutoring (6-8)	22.95%	0.48%	1.17%	18.5-55.5	4.06
District B Assigned Software (K-8)	-	-	-	27.75	5.49
District B Assigned Software (K-5)	-	-	-	27.75	6.42
District B Assigned Software (6-8)	-	-	-	27.75	2.58
District C Tutoring (K-3)	3.21%	1.84%	1.62%	30	10.17
District C Assigned Software (K-5)	1.08%	85.04%	0.52%	24	10.65
District D Interventions (K-7)	11.59%	25.23%	6.17%	15	14.48
District D Interventions (K-5)	12.19%	23.50%	5.60%	15	14.45
District D Interventions (6-7)	9.79%	30.50%	7.91%	15	14.57
District D Classroom (K-7)	-	-	-	-	9.26
District D Small Group (K-7)	-	-	-	-	11.67
District D Supervised (K-7)	-	-	-	-	12.83
District D Other (K-7)	-	-	-	-	7.07

Note. We do not display information about eligibility for interventions for which we did not receive the necessary data to capture eligibility. If intended dosage was not clearly defined at the term-level, we assumed it was 50% of the annual intended dosage.

Appendix Table A4. ELA Interventions Eligibility, Participation, and Dosage

Intervention (Grades)	Panel A: Eligible vs Treated			Panel B: Dosage	
	% eligible, not treated	% eligible, treated	% ineligible, treated	Intended dosage in hours (per term)	Average hours attended (per term)
District A Tutoring #1 (4-8)	39.17%	6.07%	0.11%	15	7.99
District A Tutoring #1 (4-5)	41.95%	9.67%	0.20%	15	10.82
District A Tutoring #1 (6-8)	37.80%	4.31%	0.07%	15	4.86
District A Tutoring #2 (K-5)	33.46%	1.56%	0.17%	15	20.80
District B Tutoring (K-5)	21.83%	2.00%	3.40%	18.5-55.5	6.52
District C Tutoring #1 (K-3)	5.97%	2.02%	1.43%	45	14.24
District C Tutoring #2 (K-2)	14.99%	3.67%	0.02%	5-37.5	9.60
District D Interventions (K-8)	12.35%	27.04%	8.28%	15	16.26
District D Interventions (K-5)	12.42%	25.28%	9.48%	15	16.07
District D Interventions (6-8)	12.20%	30.45%	5.97%	15	16.60
District D Classroom (K-8)	-	-	-	-	8.23
District D Small Group (K-8)	-	-	-	-	14.35
District D Supervised (K-8)	-	-	-	-	13.78
District D Other (K-8)	-	-	-	-	9.48

Note. We do not display information about eligibility for interventions for which we did not receive the necessary data to capture eligibility. If intended dosage was not clearly defined at the term-level, we assumed it was 50% of the annual intended dosage.

Appendix B. Intervention Descriptions

Intervention Design Choices

The interventions listed in Table 2 varied along multiple dimensions—which students were targeted; when the intervention happened; whether sessions were virtual or in-person; the qualifications and backgrounds of the providers; the student-provider ratio; and the frequency and duration of sessions. In this Appendix, we provide summaries of the variation we observed in the design of each type of program across the four districts included in this study.

Tutoring and Small Group Interventions

Each of the four districts implemented tutoring or similar small group interventions to give students additional academic support, and two districts had two distinct such interventions. Four of the six interventions provided support in math and reading, and the other two programs focused exclusively on reading. The design of these interventions varied across districts and sometimes across schools within one district. For example, one district offered both in-person and virtual tutoring programs at different schools. In two districts, these interventions were offered both during typical school hours and after school, whereas they were offered only during school hours in the other two districts. For interventions that happened during the school day, they used a "pull-out" approach that involved intervention staff removing students from core instruction, enrichment time, or advisory to work with them in a separate space. Students were targeted to receive these interventions based on a variety of at-risk factors and low academic performance (e.g., test scores, course performance).

The personnel tasked with providing these interventions varied. Two districts had only credentialed teachers providing these interventions, whereas the other two districts also employed local college students, national tutoring services, local nonprofit organizations, and/or high school students. The programs were designed for students to receive between 2–5 sessions

per week, with total expected instructional time in a given subject ranging from 30–90 hours over the course of the entire year. The planned length of intervention sessions also varied, from 20 to 90 minutes.

Out-of-School-Time Programs

Two of the four districts used Saturday programs to provide additional instruction to students. These programs were run by school sites and were not offered at all schools. They were focused on providing math and reading instruction and support (e.g., test prep, homework help) for students in K-8 in one district and K-12 in the other district. Participation was optional for both programs, but students were encouraged by teachers to participate based on their academic performance. Saturday programming was mostly offered in the spring and typically ran for 4.5 hours each day. Neither district systematically tracked Saturday programming, such that the total amount of time that enrolled students spent at Saturday programming is unknown.

Virtual Learning Tools

Virtual learning tools (e.g., iReady, ALEKS, Dreambox) were used as an academic recovery intervention in three of four districts. They used the programs to add academic time to students' days beyond core instruction. Each district targeted students who were performing below grade level or whom a teacher had recommended for participation for these interventions. One district required all students in a subset of low-performing schools to use the program. Districts assigned students to use the platform both during and outside of school hours. The amount of time students were expected to use the program ranged from 30 minutes to 3 hours per week (approximately 18 to 108 hours per year). Data on student participation and usage of these programs was available for two of the three districts.

Extended School Calendars

Finally, two of the four districts extended the school year in a subset of schools to give students additional days of instruction throughout the year. One district implemented three different models of an extended school year across participating schools, while the other district used one model. In two of the models, the additional days were not distinguishable from regular school days. In the other two models, the additional days provided slightly reduced math and reading instructional time, and more time was allocated for enrichment and social-emotional learning (SEL) activities. All extended school days were offered in person, staffed by teachers from the participating school sites, and had student–teacher ratios consistent with the schools' typical classroom ratios. The two districts varied in the schools and students they targeted, with one district targeting the lowest performing schools, and the other selecting schools based on school and family interest. Three of the programs were designed so that all students at the participating schools received additional days of instruction, whereas the remaining program specifically targeted students who were most in need of additional academic support.

The four models also varied in the total number of extra days and additional instructional time they provided to participating students relative to the traditional school calendar. One model provided students with three additional, typical days of instruction. Two other models gave students 22 additional days of instruction, but one of the models had half-days for each additional day. The full-day model consistently provided students with an additional 2 hours of reading instruction and 1.5 hours of math instruction per day, totaling an extra 44 hours of reading and 33 hours of math per year relative to the traditional calendar. The half-day model was less consistent in its instructional time across days but offered up to 1.5 hours of instruction in math and reading per day, totaling a maximum possible 33 additional hours in math and reading over the year. The remaining model provided students with up to 18 additional days of

school. For this model, each day included 1.5 hours of reading instruction and 1 hour of math instruction, totaling a maximum of 27 additional hours of reading and 18 additional hours of math over the course of the year.

Appendix C. Methods

Intervention Impacts—Variations in Estimation Models

For some districts, program participation was related not only to baseline characteristics and prior scores on the MAP Growth assessments, but also to other measures of prior achievement, such as state standardized test scores. In these cases, we estimated the following value-added model on a semester-by-semester basis, treating each semester, *t*, as a separate observation:

$$MAP_{igjts} = \alpha_0 + \alpha_1 Treatment_{igts} + \alpha_2 Eligible_{igts} + priorMAP_{igts}\gamma + priorStatetest_{igts}\beta + X_{igt}\theta + \delta_{jgts} + \epsilon_{igts}$$

This model differs from our general model described in the main text in that it includes $priorStatetest_{igts}$, a matrix with a cubic function of previous state standardized test scores in the same subject, interacted with grade level. Moreover, the availability of an additional measure of prior student performance enabled us to slightly loosen the sample restrictions for students who could be included in the estimation. Specifically, for observations of second semester program participation (i.e., interventions that took place in spring 2022), we included students so long as they had non-missing winter and spring MAP Growth test scores from 2021–22 and at least one from either a fall 2021 MAP Growth score or spring 2021 state standardized test. In models of this specification, we imputed missing prior test scores using the district-level mean and included a dummy indicating imputation, interacted with all test score variables.

sssssssssssssss

Alternative Specifications

In addition to our main VAM specification used for each district's interventions, we estimate several alternative specifications to test the sensitivity of our findings. One potential concern with our preferred specification is that value-added models that include higher order

terms of prior test scores could plausibly lead to biased estimates due to noise in the upper and lower tails of the test score distribution. While an advantage of using MAP Growth scores is that its computer adaptivity reduces noise in the tails, we estimate supplemental models that include only linear terms of prior MAP Growth scores, rather than cubic polynomial functions, as a robustness check to our results. We arrive at consistent conclusions using these models as we do with our preferred specification.

Additionally, we estimate models that use continuous measures of intervention participation (e.g. number of intervention hours attended) rather than binary indicators of intervention participation. These models assume a linear relationship between the amount of time spent in an intervention and impacts to students' test scores. We do not report the results of these models as our estimate of the hourly effects of interventions out of concern that they are likely to be endogenous-- depending on the intervention, students who are struggling more could be more likely to attend numerous hours of a recovery program. Alternatively, for more self-guided interventions such as virtual learning, more motivated students could be more likely to attend for numerous hours. Nonetheless, we find no evidence that using continuous intervention measures leads to conflicting conclusions compared to our main results.

The results of all supplemental models are available upon request.

Placebo Tests

The validity of value-added models hinges on successfully accounting for all relevant factors that differ among treated and untreated students and that could also be associated with student outcomes. One standard way to test for bias in our models is to model a treatment's impact on alternative student outcomes that should not be affected by treatment participation. Thus, for each of the math interventions we examined, we also examined impacts on reading

scores. Similarly, for reading interventions, we examined impacts on math performance. In most cases, we could not reject the hypothesis that the impact on the untargeted subject was equal to zero, as would be expected in a well- specified model. However, in some cases, such as the combined effect of participating in any math-focused interventions in District D (i.e. "All Interventions"), we estimated a positive impact on math achievement, as well as a positive impact on reading achievement. Such a pattern implies selection into the math interventions based on unobserved factors for which we could not control.

Regression Discontinuity Design

In three of the four districts in which we examined treatment effects of individual interventions, student assignment to one or more interventions was at least partially based on receiving a MAP Growth score below a certain threshold in a previous term. Assignment to treatment based on a cutoff in a continuous variable (here, MAP Growth assessment scores) often provides the opportunity to estimate treatment effects using a regression discontinuity design. We explored this methodology in all three districts, but ultimately found very weak first stages in all of them. These findings were corroborated by our district interviews, from which we learned that students' assignment to particular interventions was frequently influenced by teacher discretion, scheduling issues, and instructional capacity. We plan to continue exploring the viability of using regression discontinuity designs in future analysis, as we expect that the R2R districts' program implementation will become more systematic over time and intervention assignment may adhere more to their intended design.

Estimating Expected Effects per Hour of Intervention

To contextualize the size of the effects we estimate in this study, we also estimate the size of the effect we would expect to see for each intervention based on the number of hours students

attended the intervention if the intervention were to have the same average effect per hour as the prepandemic tutoring programs studied in Nickow et al.'s (2024) meta-analysis. The meta-analysis examines experimental evidence on tutoring programs from preschool through secondary school and finds pooled estimates of the impact of tutoring on student achievement are approximately 0.27 standard deviations in math and 0.29 standard deviations in reading. Accounting for the weights associated with each program in the meta-analysis (using Nickow et al.'s (2024) replication package), we calculate the average number of hours of tutoring offered associated with the pooled effects: 38.9 hours of math tutoring and 35.0 hours of literacy tutoring. Over 85 percent of the programs included in the study were conducted during school hours, so we then approximate student attendance in the programs as the U.S. NCES 2017-18 public school average daily attendance (ADA) rate of 93 percent (see NCES tables 203.80 and 203.20), generating estimated effects of 0.0074 standard deviations per hour of math tutoring attended and 0.0089 standard deviations per hour of literacy tutoring attended. To calculate the expected effect of each of our programs, we multiply these hourly impact estimates by the average hours (i.e., dosage) of programming that students received for each program in the study.

In Table 6, we also use these expected hourly impact estimates to approximate the hours of instruction needed to counteract the pandemic loss in each district by the following: $\frac{(Spring\ 2022 - Spring\ 2019)}{.27} * 38.9$ for math or $\frac{(Spring\ 2022 - Spring\ 2019)}{.29} * 35.0$ for reading. However, if districts face similar ongoing barriers to implementing interventions in the future, the interventions may have weaker positive effects per hour of intervention on student outcomes than the pre-pandemic tutoring programs included in Nickow et al.'s (2024) meta-analysis. Therefore, the additional hours of tutoring we estimate districts need for full recovery should be interpreted as a lower bound.