THE UNIVERSITY OF CHICAGO
EDUCATION LAB
Urban Labs

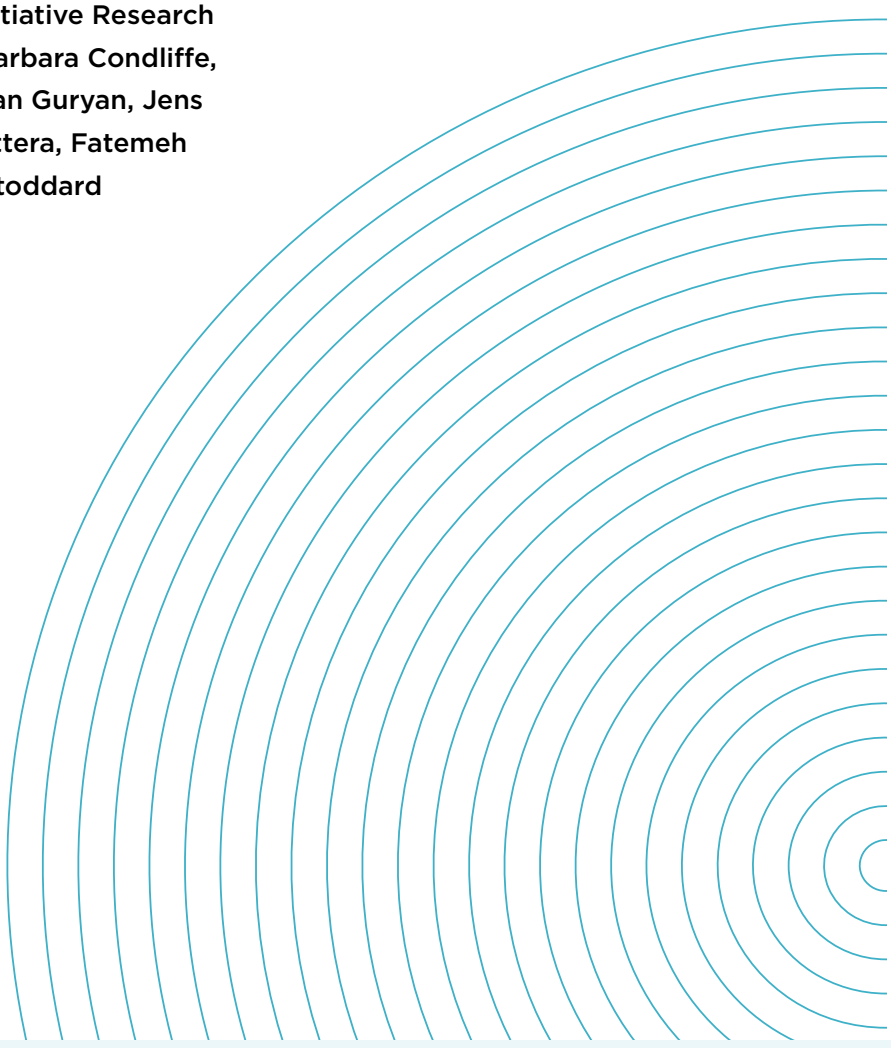
Education Insights, Real-World Impact

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

JUNE 2025

Personalized Learning Initiative Interim Report: Findings from 2023-24

Report by the Personalized Learning Initiative Research Team: Monica P. Bhatt, Terence Chau, Barbara Condliffe, Rebecca Davis, Jean Grossman, Jonathan Guryan, Jens Ludwig, Matteo Magnaricotte, Shira Mattera, Fatemeh Momeni, Philip Oreopolous, and Greg Stoddard



Acknowledgements

The Personalized Learning Initiative has been made possible through the generous support of funders including: the AbbVie Foundation; Accelerate: The National Collaborative for Accelerated Learning; Arnold Ventures; Ben and Chiara Lumpkin; the IMC Foundation; Ken Griffin, founder and CEO of Citadel and founder of Griffin Catalyst; MIT Blueprint Labs Charter School Research Collaborative; Overdeck Family Foundation; the Vivo Foundation; and the William T. Grant Foundation.

We are also deeply grateful to our school partners, including: Chicago Public Schools, Fulton County Schools, Greenville County Schools, Guilford County Schools, Miami-Dade County Public Schools, the New Mexico Public Education Department, Rocketship Public Schools, and Winston-Salem/Forsyth County Schools.

We are especially thankful to the following colleagues, whose work – across research operations, data analysis, and partner engagement – played a crucial role in bringing this project to life: Diego Aguilar, Milly Arbelo, Kevin Badon, Chris Baek, Alec Bardey, Brenda Benitez, Grace Boblick, Trayvon Braxton, Juan Castrejon, Helen Chen, Ran Cheng, Justin Cox, Lorie Chu, Axelle Clochard, Bryant Cong, Rani Corak, Hannah Dalporto, Caroline Davidson, Maria Diego Fernandez, Sonia Drohojowska, Ellen Dunn, Miriam Elkeeb, Kayla Elliott, Abigail Evans, Aidan Fitzmaurice, Reese Gaines, Victor Gamarra, Alex Gordon, Jose Guevara Hernandez, Logan Hankla, Daniel Hasak, Mervett Hefyan, Mei Huang, Maggi Ibis, Jiayu Kang, Luke Karner, Lauren Lee, Siyi Li, Anne Lombardi, Natalie Lu, Bess Markel, Bryce Marshall, Rachel McCormick, Jackie Mendez, Bhavya Mishra, Frieda Molina, William Morgan, Evelyn Morris, Michelle Ochoa, Gustie Owens, Jose Pelaez, Desiree Principe Alderson, Stephy Riega Escalante, Lauren Scarola, Sam Schneider, Laura Sikoski, Carolyn Silverman, Sadie Stockdale Jefferson, Alex Stone, Marissa Strassberger, Gargi Sundaram, Alexandria Tabasso, Julia Walsh, Weiyu Wang, Anne Warren, John Wolf, Janey Woo, and Cassie Wuest, as well as the broader Education Lab and MDRC teams.

Contents

Click to go directly to each section.

Executive summary

PRELIMINARY PLI FINDINGS FROM 2023-24 ANALYSIS:

- **Tutoring - both high dosage tutoring and sustainable high dosage tutoring - is effective overall.**
    - Pooled analyses show the effect of participating in tutoring is statistically significant and ranges from 0.06-0.09 SD, or approximately 1-2 months of additional learning. These overall effects mask considerable variability across sites.
- **Tutoring impacts seem robust across a variety of models.**
    - Lower cost models ($1200 per student) are just as effective as higher cost models ($2000 per student).
    - Virtual tutoring seems just as effective as in person tutoring in PLI sites.
- **More tutoring minutes correlate with greater learning gains.**
    - But, minutes of tutoring provided are much lower than past tutoring studies (corresponding to smaller gains in student learning).

INTRODUCTION

The dismal results from the fall 2024 release of "The Nation's Report Card" (the National Assessment of Educational Progress) confirmed what many who follow schooling outcomes for children already feared: post-pandemic academic recovery for students was middling at best, with reading and math scores still falling short of their pre-pandemic levels for every tested grade level. Even more concerning, students at the lower end of the achievement distribution lost more ground than students at the top of the distribution, exacerbating inequalities that have long persisted along race and class lines.

How could this be, given the historic investment in school funding by the federal government and the focus on instituting evidence-based practices, such as high dosage tutoring? Research suggests that the investments overall yielded significant learning per $1,000 spent, on average, which is both encouraging and falls short of the magnitude of effects needed for students to recover and thrive academically.

The cost of not doing so is catastrophic. But what about the specific impact of high dosage tutoring efforts? This report summarizes the ongoing work by our research team to understand whether and how scaling high dosage tutoring scaled in the post-pandemic environment—and what its impacts were on student achievement.

We have found both good news and more sobering news. On the one hand, tutoring works on average to significantly improve student learning above and beyond the status quo. Even more encouraging, we see positive effects for all kinds of tutoring model designs delivered in a variety of different ways across a wide range of contexts. However, overall we still see that the dosage students are getting falls far short of what would be needed to fully realize the promise of high dosage tutoring.

OVERVIEW OF THE PERSONALIZED LEARNING INITIATIVE

The insights presented in this report are derived directly from data collected to date through the Personalized Learning Initiative (PLI), a large-scale randomized controlled trial undertaken by the University of Chicago Education Lab and MDRC. Since the 2021-22 school year, the PLI research team has worked in partnership with eight sites—which include both large and small districts in urban, suburban, and rural areas, plus with an entire state, and with a charter network—randomizing more than 27,000 students to date to three different conditions:

- Evidence-based high dosage tutoring (HDT) models;
- New, innovative lower-cost tutoring models that we co-designed with our partner districts, which we called "sustainable" high dosage tutoring models (SHDT); and
- Business as usual.

Over four years, we have randomized students to one of these three conditions at the student, classroom, and sometimes grade level to capture not just whether HDT can scale successfully across the country, but also whether it is possible to lower the cost of delivering successful tutoring (in order to serve as many students as possible) by giving lower-cost SHDT to at least some students, focusing on those whom the data might suggest could still benefit from that type of help.

PLI IN 2023-24

This report focuses on our findings to date from the 2023-24 school year (see our previous report, focusing on findings from the 2022-23 school year). In 2023-24, we partnered with eight state, district, and charter education agencies nationwide, as shown below. We randomized a total of 17,330 students, and 16,435 students are in our analytic sample for 2023-24 (see Table 1 in the next section for more details).

In 2023-24, our PLI partners provided tutoring in math and reading across grades K–12, though were largely focused on early reading and middle grade math. In addition, the majority of our partners had both an HDT and an SHDT type of tutoring provided to their students.

KEY TAKEAWAYS FROM 2023-24

The good news is that tutoring can work in a post-pandemic environment across a wide range of delivery modalities (e.g., in-person tutors or virtual tutors, varying group size, across varied curricula, etc.). When benchmarked to prior studies conducted on Saga Education's high dosage tutoring models, the amount of learning *per minute* of tutoring seems similar, for the most part, even as these practitioners adapted it to their different settings—urban, rural, or suburban districts; delivered virtually or in person; and using different group sizes.

Many sites were in their pilot year in 2023-24 and began implementing tutoring in the spring; other sites had full-year implementations. In order to present readers with an "apples to apples" comparison of impacts on student learning across sites, we present student learning gains for each site and intervention per minute of tutoring.

The graphs below show the relationship between scheduled minutes in each site, or "dosage," and treatment effect, defined by student learning in standard deviation units. Figure B shows the relationship between dosage and treatment effect for HDT models and Figure C shows this same "dose-response" relationship for SHDT models. In each graph, we benchmark dosage and treatment effect findings across PLI sites with the corresponding pre-pandemic study by Saga Education ("Saga Match" for HDT in Figure B, "Saga Tech" for SHDT in Figure C).[1]

We see that dosage is considerably lower in our PLI sites in 2023-24 than in prior studies conducted by the University of Chicago Education Lab of Saga Education's tutoring programs.

For example, we see that students in the original Saga studies were scheduled for approximately 48 minutes per session daily, resulting in 2,030–4,940 minutes of tutoring received per year for students who participated. In contrast, received dosage in 2023-24 PLI site partners ranged from 631-2,287 minutes total received per year.[2]

These plots suggest the student learning per minute of tutoring is consistent across sites and studies. Consequently, increasing dosage should yield gains in student learning. While not definitive, these findings ought to encourage the field to stay the course in implementing high dosage tutoring while improving dosage to yield greater learning gains for students.

Figure A: PLI partners and designed tutoring models, 2023-24

Chicago Public Schools (IL)
- Reading for K–5, Math for 6–12
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio with edtech

Guilford County Schools (NC)
- Math for 6–8
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio with edtech

Winston-Salem/Forsyth County Schools (NC)
- Math for 6–8
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio with edtech

Rocketship Public Schools, (CA)
- Reading for K–5
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio

New Mexico (Various Schools in State)
- Math for 6–8
- Virtual tutors
- HDT: 4:1 ratio
  SHDT: NA

Fulton County Schools (GA)
- Reading for K–5, Math for 1–9
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 6 or 8:1 ratio with edtech

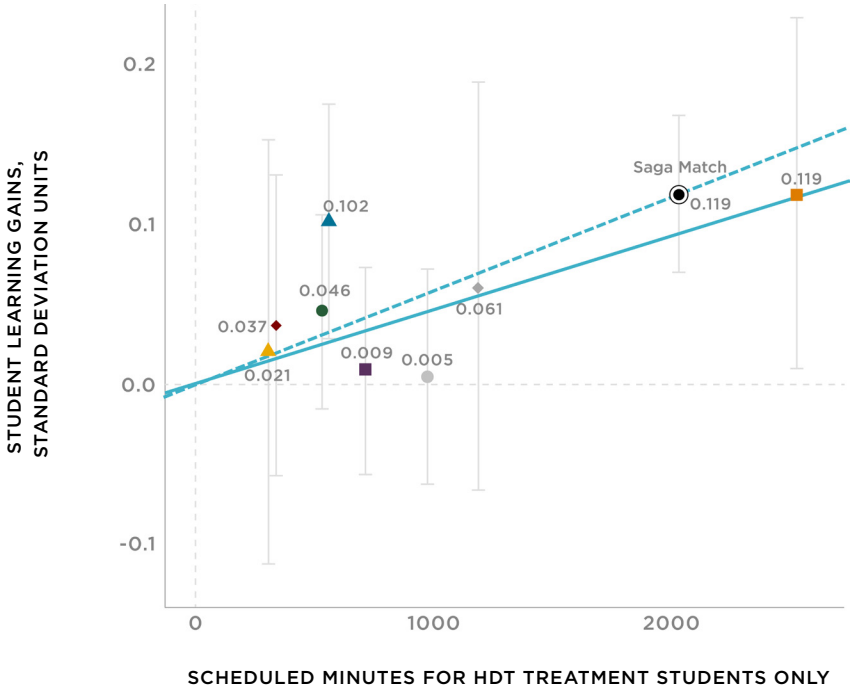
Miami-Dade County Public Schools (FL)
- Math for 6–8
- In-person tutors
- HDT: 4:1 ratio
  SHDT: NA

Greenville County Schools (SC)
- Math for 6–8
- Virtual tutors
- HDT: 3:1 ratio
  SHDT: edtech only, no tutor

[1] We use Saga Education's every day in-person 2 students:1 tutor model described in Guryan, et al. (2023) as a benchmark for HDT intervention models in PLI. We use Saga Education's in-person 4 students:1 tutor model in which students spend every other day on an education technology platform as described in Bhatt et al. (2024) as a benchmark for "sustainable" or lower-cost HDT intervention models (which we call SHDT in this report) in PLI. Both of these studies served as explicit intervention design models and starting points for districts from which they then adapted to fit their local context.

[2] Average dosage in 2023-24 was calculated by multiplying the number of sessions received (by treatment students) by the average length of a session as designed. Note that these numbers are unconditional dosage, not filtering to those who had at least one session.

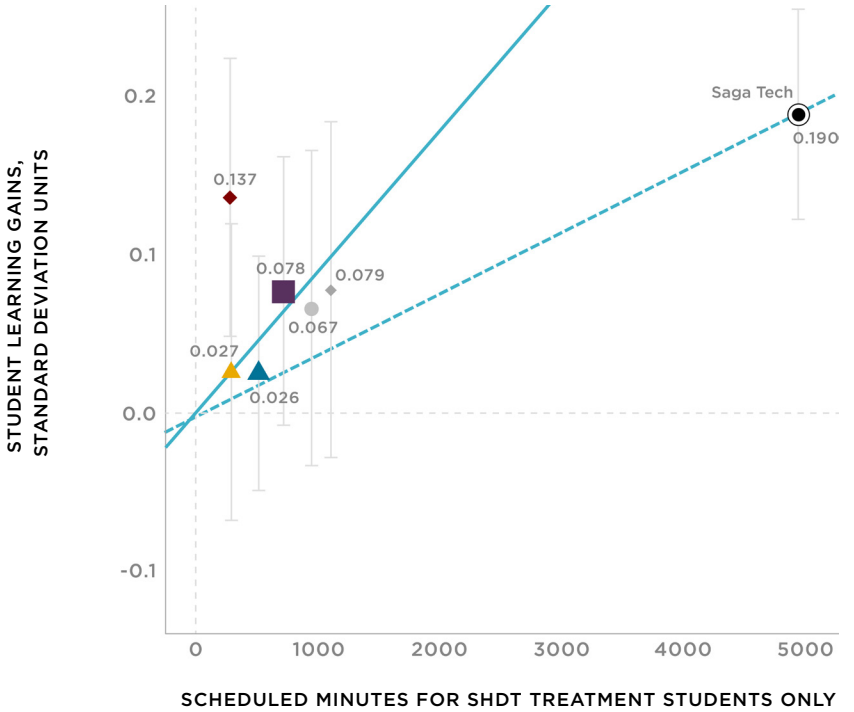
Figure B: Relationship between dosage and impact on student learning for high dosage tutoring (HDT), 2023-24



KEY TAKEAWAY

Across the eight PLI sites offering HDT in 2023-24, the relationship between minutes of tutoring and student learning gains is positive. Even more, the slope of this line is similar to that of our pre-pandemic benchmark study (Saga Match).

Figure C: Relationship between dosage and impact on student learning for sustainable high dosage tutoring (SHDT), 2023-24



KEY TAKEAWAY

Across the six PLI sites offering SHDT in 2023-24, the relationship between minutes of tutoring and student learning gains is positive. Even more, the slope of this line is similar to that of our pre-pandemic benchmark study (Saga Tech).

Note: The impact is the ITT estimate while the dosage is the minutes of tutoring scheduled for all treatment students even if they never received tutoring (i.e., the unconditional mean). Model estimates are derived from site-specific regressions. Each model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero, but do not modify their session counts. We overlay two lines passing through the origin where the slope parameter is estimated via OLS using either the ITT estimate for Saga or the available PLI ITT estimates.

Importantly, we also see that even lower-cost models are just as effective overall as higher-cost models when combined across sites. Preliminary estimates of the costs of HDT programs are about $2,000 on average and SHDT programs to be about $1,200 on average (2023 USD). The average impact on student learning of HDT pooled across sites in 2023-24 is 0.055 SD for the treatment-on-the-treated (p = 0.002) and the average impact on student learning for SHDT pooled across sites in 2023-24 is 0.085 SD for the treatment-on-the-treated (p = 0.001). This design effort to partner with school districts to find lower-cost alternatives to HDT seems like a promising approach for scaling high dosage tutoring moving forward. Ultimately, we are still interested in whether SHDT models are more effective relative to HDT models for targeting purposes—but even knowing that we can lower costs at scale and preserve efficacy is an important finding for the field (see Bhatt et al., 2024 for a more robust discussion of this approach).

The bad news, as these data and graphs suggest, is that during school year 2023-24 participating schools were unable to deliver enough tutoring during the school day to match the original studies' dosage and thus (in our estimation) its impacts. The open question is why. One possibility is that cost was a key barrier. Most site partners chose dosage targets and achieved minutes of tutoring delivered far below than what was delivered in our benchmark studies of tutoring, Saga Match and Saga Tech. But site partners fell short of even these lower intended dosage levels. Conversations with the operators suggest schools felt they simply had too many competing demands on limited instructional time, among other explanations that we detail in this brief based on data collected through interviews and surveys with tutors and school coordinators. While funding matters, our research points to a broader challenge: the many needs students have brought back to school post-pandemic have pressured schools to address multiple priorities simultaneously, making it difficult for them to carve out sufficient time for tutoring in any one subject during the school day. Early in our study, we saw many districts across the country choosing instead to move supplemental instruction to the afterschool and summer hours, but these voluntary programs attracted only a small subset of the students in need, and their instruction was often taken less seriously by students and their parents than school-day programs. (See this blog post by our study team that describes the challenges of afterschool tutoring in New Mexico, a site partner that subsequently shifted to during-the-school-day tutoring.) States and districts across the country must address this prioritization issue for HDT to reach its full potential.

Put differently, the PLI findings to date tell us that tutoring seems robust to delivery mechanism or decisions about particular program features, as long as those decisions allow for true high dosage. **Understanding how to increase dosage for students in tutoring, then, is the key challenge for the field moving forward.** One way to do this is to substantially increase investment in high dosage tutoring to saturate entire grades and schools with effective tutors. This strategy requires large amounts of consensus, political will and government investment, which may be harder to manufacture. Another strategy is to reduce costs; however, for this strategy to work, costs cannot be reduced at the expense of efficacy. A third strategy is to better target interventions to students based not just on prior performance, but on treatment responsivity—a key goal of the Personalized Learning Initiative. Finally, our school and district partners are exploring policy solutions to increase dosage such as outcomes-based tutoring and standing up data systems to track implementation monitoring. We provide examples of our own implementation monitoring reports provided to schools and districts in Appendix III, in hopes they will be useful to the field in increasing dosage.

We note here that there is a difference between a statistically significant result for research purposes and a policy decision that a district or state must make on behalf of their students. Ultimately, we think that these PLI findings from 2023-24 indicate that high dosage tutoring is still a district or state's best bet to improve student learning, given that the learning impact per minute of tutoring is largely robust to differences in tutoring models (conditional on taking place during the school day). However, the main scale challenge is to get schools to do enough tutoring. The good news is that measuring dosage is concrete and actionable, and moreover is feasible in the context of U.S. public schooling as evidenced by some of our sites, like New Mexico. The challenge is whether the field can remain focused on implementing high dosage tutoring—sustained commitment is needed to achieve the strong dosage and impacts that we know are possible for students.

WHAT COMES NEXT

The goal of PLI is not just to provide average treatment effects of tutoring models, but also to examine which tutoring interventions are most effective for different types of students. In the years to come, we plan to explore this rich heterogeneity in the hopes that pairing efficacy and cost data with specific and actionable insights into which models work best for which students will help districts better target their scarce resources to support student learning. In addition, we will pair findings from the cost study to understand benefit-cost ratios for each program model to inform district and state decision making.

In the meantime, this report shares more detail about our findings to date. The next section, "Preliminary findings from 2023-24," discusses findings to date across all partner sites, and the subsequent sections discuss the findings within each of our eight partner sites.

Preliminary findings from 2023-24

THE PERSONALIZED LEARNING INITIATIVE

The global pandemic was a once-in-a-century public health crisis, and it triggered an equally unprecedented crisis in public education. The federal government provided emergency funding to support tutoring, and U.S. Secretary of Education Miguel Cardona urged school districts nationwide to adopt tutoring as a one key strategy for addressing pandemic-related learning loss. A strong body of previous research, including studies by the University of Chicago Education Lab, has shown that tutoring is one of the most cost-effective ways to support student learning.

In 2021, we launched the Personalized Learning Initiative (PLI), a collaboration between the University of Chicago Education Lab, MDRC, and researchers from the University of Toronto and Northwestern University. Our goal: to explore whether and how the proven benefits of high dosage tutoring can be scaled to reach more students. As districts ramped up their tutoring efforts, we set out both to share the best available evidence and to work alongside school systems to generate new, practice-informed insights.

Since then, PLI has partnered with eight sites nationwide. This report, released in June 2025, highlights findings from the 2023-24 school year, during which we worked with eight diverse partner sites across the country. While our research is ongoing—including analysis to better understand what works best for which students—we are sharing our current findings now, in hopes they can offer timely, actionable insights to policymakers and education leaders.

WHY MIGHT TUTORING WORK?

A major concern for policymakers is that the pandemic will lead to not only short-term educational harm, but lifelong educational scars by reducing the benefits students get from each year of schooling in the future. Consider what classroom instruction most often looks like: 25 students or so in a classroom who vary enormously in their academic levels and hence instructional support needs. In 5th grade, for example, the average classroom has some students working at a 3rd grade level and some at an 8th grade level.[3] The school system is set up to hold teachers accountable for teaching grade-level material. Studies suggest that teachers target instruction towards roughly the 60th percentile of the distribution (Bloom, 1984). The upshot is that students who are behind grade level may have trouble engaging with grade-level instruction since that is not targeted at what they need, or "academic mismatch." The implication of academic mismatch is that the students who fall behind learn less in each subsequent year of school, and so wind up falling farther and farther behind over time.[4]

Fortunately, there is a core instructional technology known to substantially accelerate learning: high dosage tutoring (HDT). High dosage tutoring has been perhaps the most prominent, widespread response to students' learning loss resulting from the COVID pandemic, based largely on a rigorous and robust evidence base showing the consistent efficacy of high dosage tutoring across subject areas, grade levels, and educational contexts (see, e.g., Nickow, et al., 2024; Kraft, 2020; Guryan et al., 2023). In the wake of the pandemic, it is estimated that fully 66% of public schools provided high dosage tutoring to their students, despite barriers to scaling such as staffing and cost (NCES, 2024).

[3] The majority of fourth grade students have scored less than proficient on the National Assessment of Educational Progress (NAEP) math tests since 1990 (NAEP, 2022a).

[4] Empirically documenting this fact raises some subtle measurement challenges that stem from the fact that academic learning has no natural 'unit' (unlike, say, graduation rates or earnings). But the evidence is consistent with the idea that children who start performing below grade level learn less from each year of regular grade-level instruction, revealed by learning gaps that seem to widen as students progress in school (Cascio and Staiger 2012; Nielsen 2023).

However, recently released studies of tutoring in the post-pandemic landscape have yielded mixed results (see, e.g., Robinson et al., 2024; Kraft, Edwards and Cannata, 2024). PLI is premised on the idea that to fully realize the promise of tutoring, we must better understand the efficacy of these interventions in the post-pandemic context; develop and generate knowledge about lower-cost, more sustainable versions of tutoring; and explore variation of these impacts by student and program characteristics in order to improve targeting.

In nationwide surveys, teachers have reported the hardest parts of classroom teaching are i) personalizing instruction and ii) classroom management.[5]

By reducing student-instructor ratios, HDT makes it easier for tutors to personalize instruction, increasing the time on tasks academically relevant for each student. In addition, smaller student-instructor ratios help strengthen relationships between the instructor (teacher or tutor) and student so there can be improved student engagement through a "social capital" or mentoring effect (Coleman, 1988; Herrera et al., 2011). We visualize this theory of change in Figure D below, capturing these mechanisms for student engagement through mediating and measurable factors such as participation and dosage in the context of our study.

Figure D: Theory of Change

Target Population

District-targeted **K-12 students**

↓

Intervention: Key Components

- Delivered **during the school day**
- High dosage: target of three 30 minute sessions per week, totalling **90 minutes weekly**
- Uses a **structured curriculum**, ideally one aligned with the district core instructional curriculum
- **Target ratios** of tutors to students:
  - HDT: target of 4 students to 1 tutor
  - SHDT: target of 8 students to 1 tutor, and/or integration of adaptive edtech to personalize instruction for students (either as a substitute or complement to tutor-delivered instruction)

↓

Mediators

- **Student engagement**, as measured by:
  - Takeup (defined as attending at least one session) and
  - Dosage (defined as the number of sessions per student per week; or number of minutes per student per week)

↓

End of Year Outcomes

- Primary outcome:
  - **Increased learning** as measured by standardized test scores
- Secondary outcomes, including:
  - Increased school attendance
  - Increased grade point average (GPA)
  - Fewer course failures

[5] For example, in the School and Staffing Survey (SASS), 43% of new elementary school teachers and 47% of new secondary school teachers say they felt not at all or only somewhat prepared to deal with classroom management; 41% of new elementary school teachers and 44% of new secondary-school teachers said they were unprepared or only somewhat prepared to differentiate instruction. (From original author tabulations of SASS data).

HOW DID DISTRICT AND STATE PARTNERS TAKE THIS EVIDENCE BASE AND APPLY IT TO THEIR OWN CONTEXTS?

In each of our PLI sites, we partnered with policymakers and district and school practitioners to co-design how to operationalize the tenets of HDT and SHDT to fit their local context. Figure E below shows the program features of these co-designed models across our sites.

While there are commonalities with respect to student-tutor ratios and scheduled time during the school day, there is considerable variation in subject area, modality (in-person or virtual), grade level, inclusion of edtech, etc. Future PLI analysis will model how this heterogeneity affects the impact of the program to provide actionable information to policymakers and practitioners.

WHO WAS PART OF PLI IN 2023-24?

Table 1 below shows the sample of students, by site, that were randomized by the study in the 2023-2024 school year.[6] The randomization design in each site is also shown.[7]

The 16,435 students in the total analytic sample represent a broad swath of backgrounds and identities. The group includes larger numbers of Hispanic, Black, and white students, as well as smaller numbers of Asian, Native American, Pacific Islander, and multiracial students.

The vast majority of enrolled students receive free and reduced lunch (more than 80% of the HDT group and more than 90% of the SHDT group). Roughly 15% are diverse learners, and between 24-30% speak English as a second language.

Figure E: Tutoring models as designed, 2023-24

Chicago Public Schools (IL)
- Reading for K–5, Math for 6–12
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio with edtech

Guilford County Schools (NC)
- Math for 6–8
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio with edtech

Winston-Salem/Forsyth County Schools (NC)
- Math for 6–8
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio with edtech

Rocketship Public Schools, (CA)
- Reading for K–5
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 8:1 ratio

New Mexico (Various Schools in State)
- Math for 6–8
- Virtual tutors
- HDT: 4:1 ratio
  SHDT: NA

Fulton County Schools (GA)
- Reading for K–5, Math for 1–9
- In-person tutors
- HDT: 4:1 ratio
  SHDT: 6 or 8:1 ratio with edtech

Miami-Dade County Public Schools (FL)
- Math for 6–8
- In-person tutors
- HDT: 4:1 ratio
  SHDT: NA

Greenville County Schools (SC)
- Math for 6–8
- Virtual tutors
- HDT: 3:1 ratio
  SHDT: edtech only, no tutor

[6] PLI took place in the 2021-22, 2022-23, and 2023-24 school years and is still enrolling sample members.

[7] If the level of randomization differs across schools within a site, all levels are listed.

Table 1: Randomization in 2023-24

	Number of Students Randomized to Analytic Sample					Research Design	
	HDT Sample		SHDT Sample		Total		
	HDT	BAU	SHDT	BAU	Total	Randomization Design	Randomization Level
Chicago	1484	1040	596	438	3558	2-arm	Student
Fulton	779	1451	1037	1324	3267	2-arm, 3-arm	Student, Classroom, Teacher
Greenville	665	670	674	670	2009	3-arm	Student
Guilford	393	646	376	646	1415	3-arm	Classroom
Miami	980	1291			2271	2-arm, 3-arm	Classroom
New Mexico	572	890			1462	2-arm	Grade, Student
Rocketship	302	303	279	303	884	3-arm	Grade
Winston-Salem	497	624	448	624	1569	3-arm	Classroom
Total	5672	6915	3410	4005	16435		

Note: As a reminder, HDT stands for high dosage tutoring, SHDT stands for sustainable high dosage tutoring, and BAU stands for business as usual (our control group). This table only includes randomized students that are included in our study.[8] Some students that were randomized were not included in our study. For sites with a 3-arm design, the total number of students is not equal to the HDT sample plus the SHDT sample because some BAU students are in both samples (Fulton = 1,324, Greenville = 670, Guilford = 646, Miami = 44, Rocketship = 303, Winston-Salem = 624).

[8] Table 1 only includes randomized students that are included in our study. Not all students that were randomized were included in our study. Per the pre-analysis plan, we exclude from the analysis any randomization blocks where (under clustered randomization) control clusters were offered tutoring while no treatment cluster was offered tutoring (1 randomization block with 88 students from Fulton, 1 school with 71 students from New Mexico, and 3 randomization blocks with 135 students from Miami). We exclude randomization blocks with no control students (12 randomization blocks with 418 students from Fulton where students were randomly assigned to either HDT or SHDT). We also exclude students who were in multiple randomization blocks at the time of randomization, which occurred in classroom-level randomization (3 students from Miami, 2 from Winston-Salem). Additionally, we exclude randomization units that were ineligible for the study and were only randomized because we had incorrect information about them at the time of randomization. For this reason, we exclude students for which we had the wrong grade information at the time of randomization (5 students from Fulton) and classes that were too advanced to be eligible for randomization (3 classes with 58 students from Miami). After excluding those ineligible classes from Miami, 2 randomization blocks were left with only control students, so the rest of those blocks are excluded as well (81 students from Miami).

Table 2: Baseline balance, pooled analysis sample 2023-24, HDT

Covariate	Control Mean N = 6177	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	11.32	-0.06	0.361	0.461	11033
% Male	0.5	0	0.984	0.982	11021
% Hispanic (Race or Ethnicity)	0.5	0	0.663	0.672	11027
% Black	0.39	-0.01	0.166	0.216	11023
% White	0.31	0	0.821	0.857	11023
% Asian	0.02	0	0.193	0.325	11023
% Native American	0.03	0.01	0.012**	0.015**	11023
% Pacific Islander	0.01	0	0.571	0.626	9787
% Multiracial	0.03	0	0.619	0.641	5423
% English as a Second Language	0.28	0	0.993	0.993	10965
% Receiving Free/Reduced Lunch	0.83	0	0.777	0.795	7137
% Diverse Learner	0.17	-0.01	0.216	0.293	9113
% Homeless	0.04	0	0.581	0.606	5040
Number of Days Attended	152.7	0.78	0.356	0.397	5951
Overall GPA	2.76	0.02	0.675	0.722	6791
Overall GPA x Math	2.26	0.01	0.711	0.768	6791
Overall GPA x ELA	0.49	0	0.65	0.645	6791
Latest Available Math Score	0.02	0.06	0.12	0.193	9731
Second Latest Available Math Score	-0.01	0.03	0.482	0.539	8121
Latest Available Reading Score	0	0.03	0.339	0.417	9468
Second Latest Available Reading Score	0.02	0.02	0.59	0.643	9030
***F*-Test - Baseline Cov.**			0.082*	0.925	11109

Note: The analysis sample is composed of HDT and BAU students randomized in a block in the 2023-24 school year in any site who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N." The latest and second-latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test. Grade indicators are not shown for brevity.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g., HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: **** $p < 0.01$*, *** $p < 0.05$*, ** $p < 0.10$*.*

Table 3: Baseline balance, pooled analysis sample 2023-24, SHDT

Covariate	Control Mean N = 3677	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	11.53	0.02	0.836	0.85	6620
% Male	0.5	-0.01	0.638	0.68	6610
% Hispanic (Race or Ethnicity)	0.39	0.01	0.448	0.544	6610
% Black	0.5	-0.03	0.051*	0.088*	6610
% White	0.12	0.01	0.476	0.566	6610
% Asian	0.03	0.01	0.109	0.202	6610
% Native American	0.01	0	0.177	0.228	6610
% Pacific Islander	0	0	0.505	0.532	5367
% Multiracial	0.04	0.01	0.271	0.277	4155
% English as a Second Language	0.25	0.01	0.5	0.62	6610
% Receiving Free/Reduced Lunch	0.92	-0.01	0.164	0.289	3406
% Diverse Learner	0.16	-0.01	0.332	0.41	4647
% Homeless	0.03	0.01	0.21	0.237	2881
Number of Days Attended	153.96	-0.33	0.717	0.744	4856
Overall GPA	2.66	0.08	0.19	0.285	4013
Overall GPA x Math	2.49	0.08	0.178	0.268	4013
Overall GPA x ELA	0.18	0	0.407	0.389	4013
Latest Available Math Score	0.01	0.11	0.01**	0.026**	6340
Second Latest Available Math Score	0.01	0.05	0.372	0.471	4884
Latest Available Reading Score	0.01	0.07	0.154	0.232	5303
Second Latest Available Reading Score	0.02	0.05	0.289	0.372	5048
F-Test - Baseline Cov.			0.116	0.913	6668

Note: The analysis sample is composed of SHDT and BAU students randomized in a block in the 2023-24 school year in any site who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with SHDT students are included. Reported p*-values test the difference in means for the SHDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N." The latest and second-latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test. Grade indicators are not shown for brevity.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows:*
*** p < 0.01, ** p < 0.05, * p < 0.10.*

HOW MUCH TUTORING DID STUDENTS RECEIVE?

We have near perfect visibility into whether or not students assigned to tutoring received any tutoring and, if so, how much tutoring they got. In PLI sites, 86% of students assigned to HDT and 79% of students assigned to SHDT received at least one session. These are high take-up rates compared to prior interventions of tutoring (see, e.g., Guryan et al., 2023). Randomization status was also preserved as we see low control crossover overall—less than 3% of students assigned to the control group received tutoring services, though they did receive all other services a school had to offer.

Students who participated in tutoring received on average 29 sessions in HDT, for a total of over 17 hours of tutoring during the 2023-24 school year. On average, students assigned to receive SHDT who participated in at least one session attended 20 sessions and just over 13 hours of tutoring for the school year. This dosage is considerably lower than the 34-82 hours we estimate students received in our benchmark pre-pandemic studies with Saga Education (see, e.g., Guryan et al., 2023 and Bhatt et al., 2024).

Table 4: Take-up and dosage, pooled analysis sample 2023-24

HDT Analysis Sample		
	BAU	HDT
N	6177	4932
% Received Treatment	2.75	86.03
Average Attended Sessions (Conditional)	29.49	29.31
Average Attended Sessions (Unconditional)	0.82	25.22
Average Attended Minutes (Conditional)	927.21	1029.82
Average Attended Minutes (Unconditional)	25.87	885.95
% Students Missing Dosage Data	0	0

SHDT Analysis Sample		
	BAU	SHDT
N	3677	2991
% Received Treatment	2.91	79.07
Average Attended Sessions (Conditional)	6.21	20.31
Average Attended Sessions (Unconditional)	0.2	16.06
Average Attended Minutes (Conditional)	207.52	801.17
Average Attended Minutes (Unconditional)	6.60	631.16
% Students Missing Dosage Data	0	0

Note: The analysis sample is the sample of randomized students (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

WHAT WERE THE EFFECTS OF THESE TUTORING PROGRAMS ON STUDENT LEARNING?

Our 2023-24 analysis found that tutoring can be effective across a wide range of implementation models—both HDT and SHDT—and formats ranging from in-person tutors to virtual tutors. HDT models, which cost on average $2,000 show overall intent-to-treat treatment effect sizes of 0.046 (*p*-value = 0.015; TOT = 0.055 with *p*-value = 0.002), equivalent to approximately one month of learning (following Hill, et al., 2008).[9] SHDT models, which cost on average $1,200 show overall treatment effect sizes of 0.065 (*p*-value = 0.001; TOT = 0.085; *p*-value = 0.001), or approximately two months of learning. The four figures below show the impact—both intent-to-treat and treatment-on-the-treated—of HDT and SHDT, pooled across sites and for each site.

These overall average effects belie considerable variation across sites and also in program design. As shown in Figure G, the TOT treatment effects for HDT for students who participated across sites ranged from 0.008 SD to 0.13 SD, at varying levels of statistical significance, as noted by the confidence intervals around each treatment effect. (For instance, the impact of tutoring on student learning in New Mexico in 2023-24 could be as high as 0.25 SD or as low as 0.01 SD.) As shown in Figure I, the TOT treatment effects of SHDT ranged from 0.034 SD to 0.16 SD, again with varying levels of significance. These site-specific estimates exhibit wide variation within site, including sites where multiple tutoring models were implemented; however, when pooled overall we see the effects are positive and statistically significant, despite variation in program models.

The range of treatment effects in our studies is considerably smaller than the average treatment effects from pre-pandemic pre-ESSER funding meta-analyses of tutoring studies (see, e.g., Nickow, Oreopolous & Quan, 2021). These PLI impact estimates are on par with other studies that have found smaller effects on high dosage tutoring in the post-pandemic ESSER period as well and with other educational interventions implemented at scale (see, e.g., Kraft 2024, Robinson, 2024).

One explanation of this phenomenon may be that the "business as usual" counterfactual condition, which students who were not randomly assigned to the PLI tutoring models received, was likely rich with other personalized learning efforts. In Spring 2024, 80% of school leaders from a nationally representative survey of school leaders reported stable or increased demand for evidence-based interventions like tutoring. Our own surveys of school coordinators show a high degree of individualized instructional interventions were available to all students in PLI study schools—including for students in the control group. For example, over 80% of PLI school coordinators reported that more than 75% of their student population has access to computer assisted learning platforms for math and/or literacy. As such, the PLI estimates of the effects of tutoring are on top of this rich counterfactual condition, which will likely no longer be in place as the ESSER funding ends. In other words, the effects of the same tutoring models in the absence of ESSER funding may be larger, as students return to receiving less personalized instruction in their usual school experiences. This is consistent with estimates that every $1,000 invested in schools from ESSER spending led to a 0.002-0.007 SD increase in student learning (Dewey et al., 2024).

While these estimates should be considered preliminary, they signal that tutoring implemented at scale is still a strategy that increases student learning on average, above and beyond what schools were already doing in the context of ESSER and post-pandemic catch up supports. Additionally, it is notable that there are lower-cost versions that can be implemented and be equally effective as higher-cost versions.

[9] To convert estimates from standard deviations to months of learning, we start by estimating the expected yearly growth for our sample. This is done by averaging the expected growth expressed in standard deviations estimated by Hill et al. (2008) for each student in the analysis sample based on their grade and whether they were assigned to math or reading tutoring. The expected growth for the analysis sample expressed in standard deviations is 0.47 SD for the HDT sample and 0.43 for the SHDT sample. We then convert the estimated impact of tutoring into years of learning and multiply the resulting number by 9 to express it in months of learning.

Figure F: Effect of being assigned to HDT on student learning (ITT), 2023-24

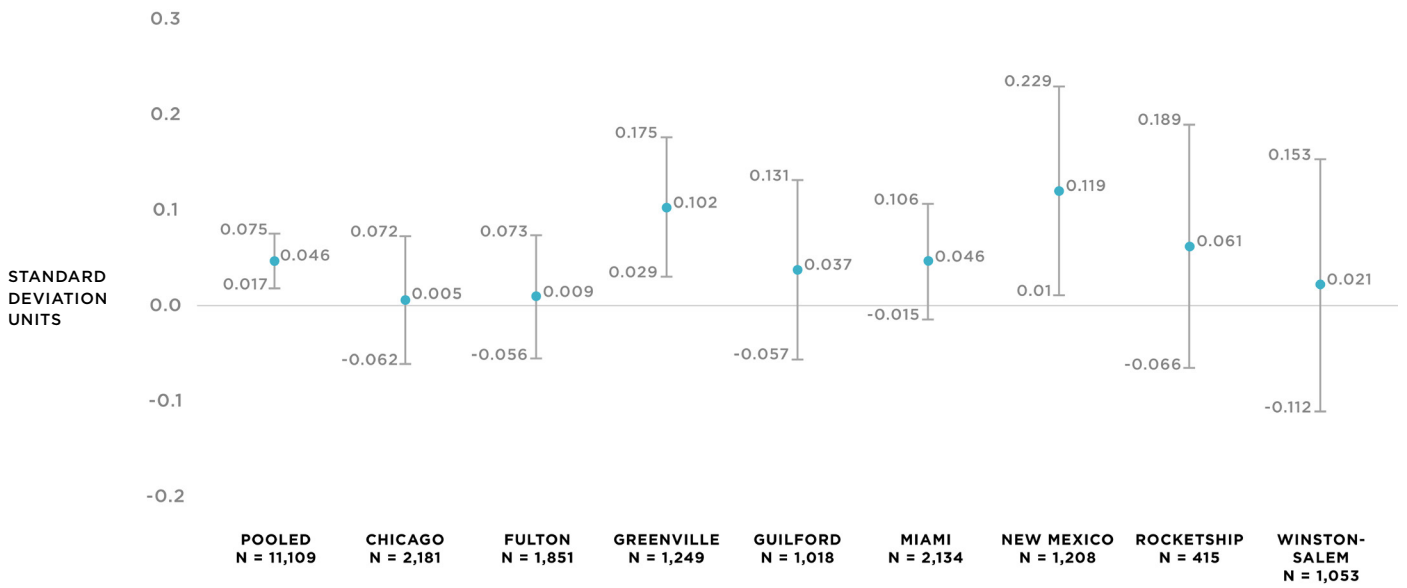
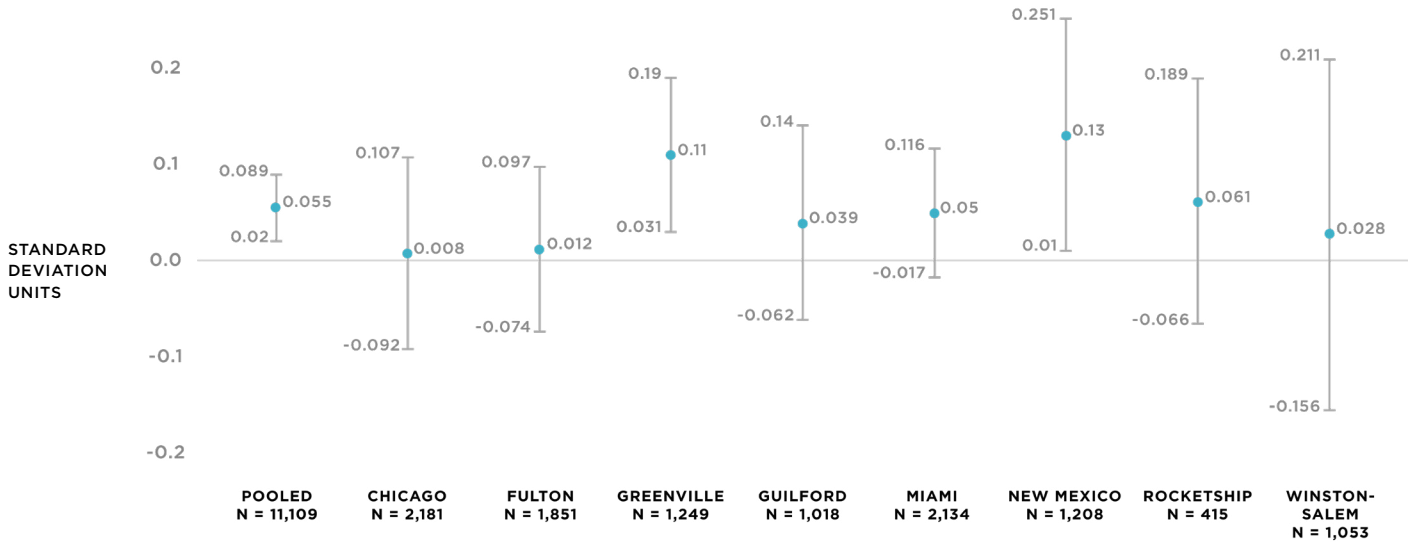


Figure G: Effect of participating in HDT on student learning (TOT), 2023-24



Note: Model estimates are derived from site-specific regressions. Each model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

Figure H: Effect of being assigned to SHDT on student learning (ITT), 2023-24

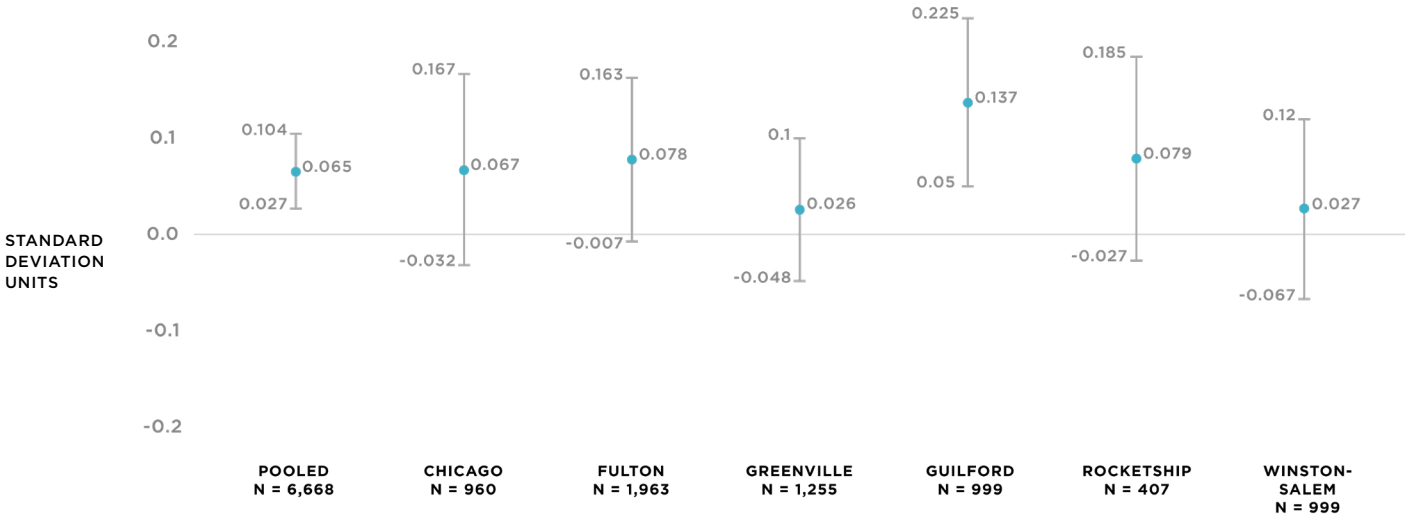
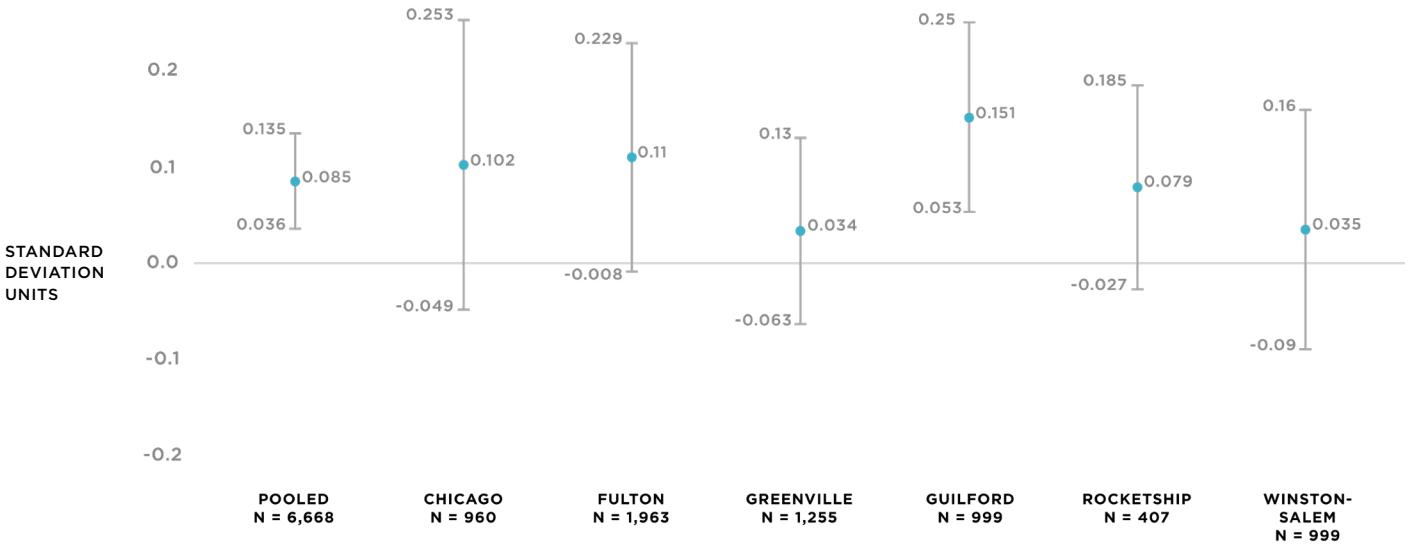


Figure I: Effect of participating in SHDT on student learning (TOT), 2023-24



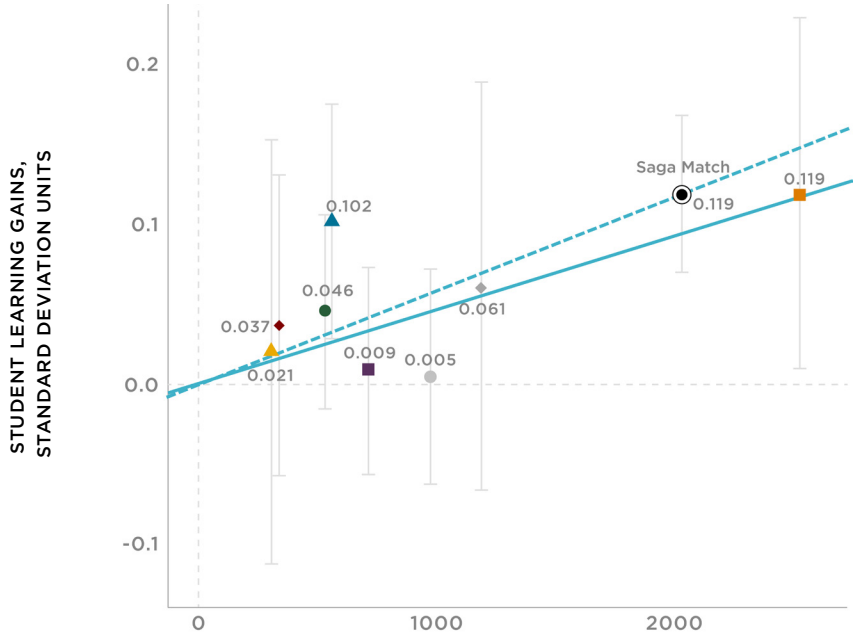
Note: Model estimates are derived from site-specific regressions. Each model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

WHAT IS THE RELATIONSHIP BETWEEN DOSAGE AND STUDENT LEARNING?

Many sites were in their pilot year in 2023-24 and began implementation in the spring; other sites had full-year implementations. In order to present readers with an "apples to apples" comparison of impacts on student learning across sites, we present student learning gains for each site and intervention per minute of tutoring. The graphs below show the relationship between scheduled minutes in each site, or "dosage," and treatment effect, defined by student learning in standard deviation units. Figure J shows the relationship between dosage and treatment effect for HDT models and Figure K shows this "dose-response" relationship for SHDT models. In each graph, we benchmark dosage and treatment effect findings across PLI sites with the corresponding pre-pandemic study by Saga Education study ("Saga Match" for HDT in Figure J, "Saga Tech" for SHDT in Figure K).[10]

Overall, we see that dosage is considerably lower in our PLI sites than in prior "benchmark" studies conducted by the University of Chicago Education Lab. For example, Education Lab's evaluation of Saga Education's tutoring program, which found effect sizes of 0.26 SD on average, had approximately 48 scheduled minutes per session, resulting in 2,030 – 4,940 minutes (roughly 34 – 82 hours) of tutoring received per year. In contrast, while the range of scheduled minutes per session in PLI sites ranged from 30-80 minutes per session, the actual dosage received ranged from 631-2,287 minutes (roughly 10 – 38 hours) total received by a student per year.[11]

Figure J: Relationship between dosage and impact on student learning for high dosage tutoring (HDT), 2023-24



KEY TAKEAWAY

Across the eight PLI sites offering HDT in 2023-24, the relationship between minutes of tutoring and student learning gains is positive. Even more, the slope of this line is similar to that of our pre-pandemic benchmark study (Saga Match).

- Chicago
- Fulton
- Greenville
- Guilford
- Miami
- New Mexico
- Rocketship
- Winston Salem
- Relationship between dosage and impacts: **PLI**
- Relationship between dosage and impacts: **Saga Match**

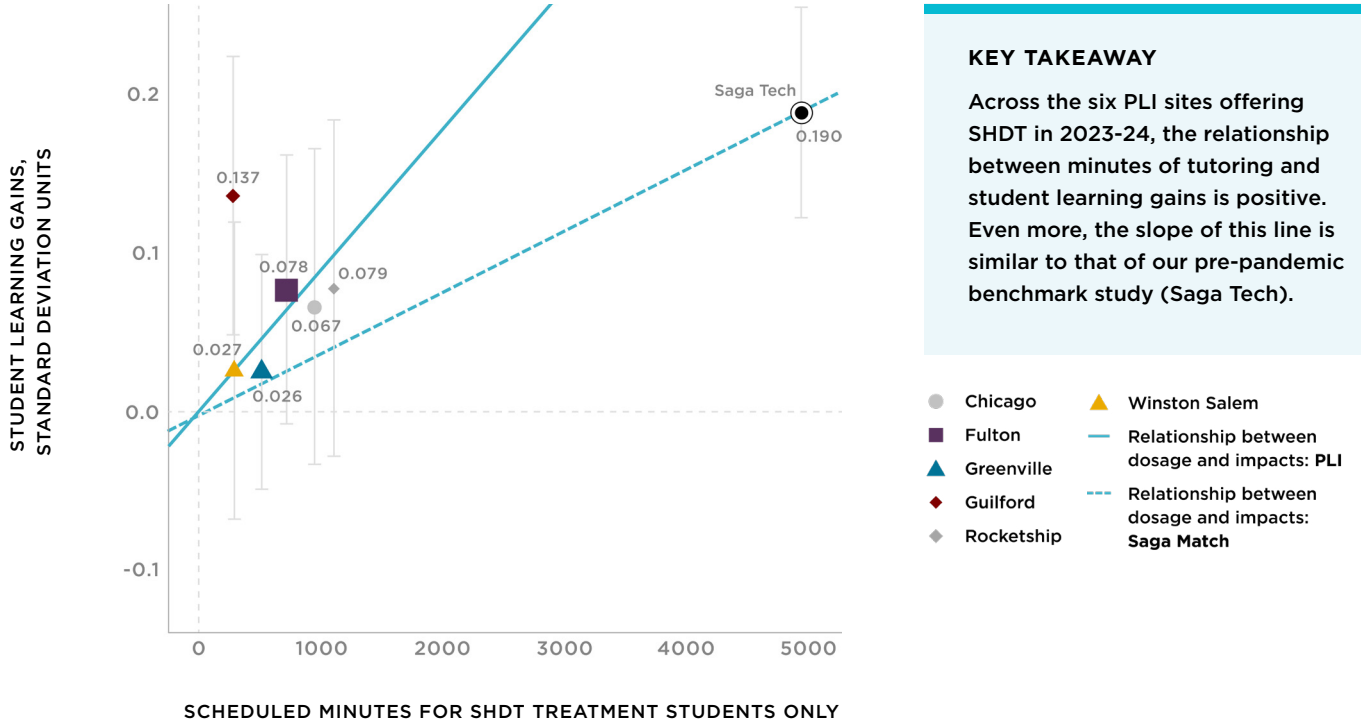
Note: The impact is the ITT estimate while the dosage is the minutes of tutoring scheduled for all treatment students even if they never received tutoring (i.e., the unconditional mean). Model estimates are derived from site-specific regressions. Each model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero, but do not modify their session counts. We overlay two lines passing through the origin where the slope parameter is estimated via OLS using either the ITT estimate for Saga or the available PLI ITT estimates.

[10] We use Saga Education's every day in-person 2 students: 1 tutor model described in Guryan, et al. (2023) as a benchmark for HDT intervention models in PLI. We use Saga Education's in-person 4 students:1 tutor model in which students spend every other day on an education technology platform as described in Bhatt et al. (2024) as a benchmark for "sustainable" or lower-cost HDT intervention models (which we call SHDT in this report) in PLI. Both of these studies served as explicit intervention design models and starting points for districts from which they then adapted to fit their local context.

[11] Average dosage in 2023-24 was calculated by multiplying the number of sessions received (by treatment students) by the average length of a session as designed using the above numbers. Note that these numbers are unconditional dosage, not filtering to those who had at least one session.

Figure K: Relationship between dosage and impact on student learning for sustainable high dosage tutoring (SHDT), 2023-24



KEY TAKEAWAY

Across the six PLI sites offering SHDT in 2023-24, the relationship between minutes of tutoring and student learning gains is positive. Even more, the slope of this line is similar to that of our pre-pandemic benchmark study (Saga Tech).

Note: Model estimates are derived from site-specific regressions. Each model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

Dosage is calculated as of the latest assessment in the primary index outcome. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero, but do not modify their session counts. We overlay two lines passing through the origin where the slope parameter is estimated via OLS using either the ITT estimate for Saga or the available PLI ITT estimates.

These plots suggest the positive relationship between dosage and student learning holds such that a focus on increasing dosage would yield gains in student learning, by definition. While not definitive, these findings ought to encourage the field to stay the course in focusing on how to improve implementation of high dosage tutoring to yield greater learning gains for students.

WHAT COMES NEXT?

As mentioned above, PLI is a multiyear study, with randomization in the years prior to 2023-24, and continuing since. After sample enrollment and program implementation is complete in 2026, future reports will present findings about the average impact across all years and sites, as well as explore how these impacts vary by student characteristics, program features, and context. This interim report presents the impact of tutoring on this sample of students in the 2023-2024 school year, with the subsequent sections focusing on site-by-site impacts.

Preliminary findings from Chicago tutoring

THE PARTNERSHIP

Since 2021, the Personalized Learning Initiative (PLI) research team has partnered with Chicago Public Schools (CPS) to study the impacts of tutoring as provided by the CPS Tutor Corps. In 2023-24, we evaluated two types of tutoring in 33 CPS schools to explore the impacts on student learning.

THE INTERVENTIONS

CPS has their own in-house group of tutors, the CPS Tutor Corps, who were deployed at over 200 of the more than 600 CPS schools. Two vendors—Saga Education and Amplify—provided technical assistance to CPS Tutor Corps in its math and reading tutoring, respectively. CPS decided that tutors in grades K–5 would support reading, while tutors in grades 6–12 would provide support in math. The PLI study partnered with 33 elementary and high schools to study two types of tutoring—HDT and SHDT—provided by the CPS Tutor Corps.

- **HDT - 4:1 tutoring** – At 23 of our study schools, tutors worked with groups of four students. Students were supposed to receive three 30-minute sessions of tutoring per week. HDT was piloted in both K–5 (for reading) and 6–12 schools (for math).
- **SHDT - 8:1 tutoring with edtech** – Another 10 study schools participated in a pilot, where tutors were assigned to groups of eight, rather than four, students. In each SHDT tutoring session, four out of the eight students in a tutoring group worked directly with the tutor, while the other four worked independently using edtech. In the next session, the roles were reversed: the students who had used edtech worked with the tutor, and those who had worked with the tutor switched to edtech. SHDT's target dosage was also slightly higher than the target dosage for HDT, with four 30-minute sessions per week. SHDT was piloted only in grades 6–12 (and thus, was tested only for math and not reading).

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** – In all schools, there were more students eligible for tutoring than the school had the capacity to serve. So, we conducted a fair lottery to assign students to the available tutoring slots (either HDT or SHDT, depending on the school). Because of this randomization, differences in end of year test scores between the control and treatment groups were attributable directly to the interventions.
- **Surveys with tutors** – The PLI team obtained a roster of all tutors in PLI study schools in CPS in fall of 2023 (N = 98 tutors) and again in the spring of 2024 (N = 82 tutors). Response rates were high, with 94% responding to the fall survey and 85% responding to the spring survey.
- **Surveys and interviews with coordinators** – A school staff member responsible for coordinating the tutoring program in each of the PLI study schools was surveyed in spring of 2024 (N = 33 school staff) and 85% responded. In addition, we conducted interviews with a subsample of 13 of these school coordinators. These interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation and resource use.

THE STUDENTS

School staff at the 33 schools determined which students were eligible for the tutoring interventions. Overall, 3,558 students were randomized in 2023-24 across our 33 schools – 2,524 in the HDT evaluation (treatment and control) and 1,034 in the SHDT evaluation (treatment or control). Our analysis sample includes 3,141 students for whom we observe end-of-year test scores in the tutoring subject area.

In our HDT analysis sample, approximately 97% of the study participants were students of color and around 40% identified as English as a second language. In comparison, the SHDT analysis sample included about 98% students of color and roughly 19% with English as a second language.

As shown in the tables below, while some individual observable baseline characteristics exhibit statistically significant imbalances between the treatment and comparison groups, both analysis samples are overall well balanced, as confirmed by the F-statistics.

Table 5: Baseline balance, Chicago 2023-24, HDT

Covariate	Control Mean N = 898	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	10.9	-0.01	0.381	0.392	2162
% Male	0.46	0.01	0.754	0.773	2162
% Black	0.39	-0.02	0.153	0.185	2171
% Hispanic	0.59	0.03	0.082*	0.091*	2171
% Other Race	0.02	-0.01	0.395	0.383	2171
% English as a Second Language	0.41	0.02	0.318	0.315	2173
% Receiving Free/Reduced Lunch	0.89	-0.01	0.468	0.446	2173
% Diverse Learner	0.12	-0.02	0.223	0.236	2173
% Homeless	0.06	0.01	0.44	0.435	2102
Number of Days Attended	151.06	-0.02	0.988	0.988	2099
Overall GPA	3.18	0.01	0.707	0.707	1992
Overall GPA x Math	1.48	0	0.912	0.917	1992
Overall GPA x ELA	1.7	0.01	0.514	0.505	1992
Latest Available Math Score	-0.05	0.03	0.58	0.552	1259
Second Latest Available Math Score	-0.05	-0.02	0.716	0.744	1237
Latest Available Reading Score	-0.01	-0.04	0.355	0.355	1950
Second Latest Available Reading Score	0	-0.05	0.348	0.349	1905
% Grade 1	0.07	0	0.431	0.661	2181
% Grade 2	0.1	0	0.335	0.591	2181
% Grade 3	0.16	0	0.291	0.45	2181
% Grade 8	0.14	0	0.247	0.478	2181
% Grade 9	0.02	0	0.247	0.478	2181
F-Test - Baseline Cov.			0.811	0.917	2181

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in Chicago who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and control groups. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test. For brevity, we only show grade-level indicators that vary within randomization blocks.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly reassign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g., HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: **** $p < 0.01$*, *** $p < 0.05$*, ** $p < 0.10$*.*

Table 6: Baseline balance, Chicago 2023-24, SHDT

Covariate	Control Mean N = 409	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	15.07	-0.01	0.718	0.717	960
% Male	0.42	0.06	0.125	0.144	960
% Black	0.55	0.02	0.487	0.513	960
% Hispanic	0.43	-0.01	0.774	0.794	960
% Other Race	0.02	-0.01	0.338	0.359	960
% English as a Second Language	0.2	0.03	0.212	0.218	960
% Receiving Free/Reduced Lunch	0.87	-0.07	0.006***	0.009***	960
% Diverse Learner	0.21	-0.01	0.74	0.752	960
% Homeless	0.05	0.01	0.477	0.496	918
Number of Days Attended	157.33	-0.41	0.794	0.807	918
Overall GPA	2.91	0.08	0.063*	0.069*	858
Overall GPA x Math	2.91	0.08	0.063*	0.069*	858
Latest Available Math Score	0.11	-0.01	0.838	0.827	933
Second Latest Available Math Score	0.03	0.05	0.55	0.528	927
Latest Available Reading Score	0.12	0.02	0.756	0.776	932
Second Latest Available Reading Score	0.04	0.03	0.69	0.709	926
***F*-Test - Baseline Cov.**			0.228	0.288	960

Note: The analysis sample is composed of SHDT and BAU students randomized in a block in 2023-24 in Chicago who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with SHDT students are included. Reported p-values test the difference in means for the SHDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations varies reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting F-statistic from this regression. Missingness indicators are included in the regression model but not in the F-test. Grade indicators are not shown for brevity.

We report both the p-value and randomization inference p-values, to avoid distributional assumptions. To calculate the randomization inference p-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and estimate the corresponding test-statistic (p-value for pairwise tests, p-value of the F-statistic for joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI p-value. p-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows:
**** p < 0.01, ** p < 0.05, * p < 0.10.*

TAKE-UP AND DOSAGE

Our school partners largely adhered to the randomized assignments. In the HDT sample, approximately 80% of students assigned to the treatment group received at least one tutoring session. We also observed that 11% of students in the control group received some amount of tutoring. In our SHDT analysis sample, 65% of treatment students and 3% of control students ended up receiving at least one session of tutoring.

Among students in the analysis sample who had at least one session of tutoring, the HDT students received an average of 43 tutoring sessions over the year, while the SHDT students who had received at least one session received an average of 37 sessions over the course of the year.

Table 7: Take-up and dosage, Chicago 2023-24

HDT Analysis Sample		
	BAU	HDT
N	898	1,283
% Received Treatment	11.14	79.50
Average Attended Sessions (Conditional)	40.48	43.23
Average Attended Sessions (Unconditional)	4.51	34.37
% Students Missing Dosage Data	0.00	0.00

SHDT Analysis Sample		
	BAU	SHDT
N	409	551
% Received Treatment	3.42	64.79
Average Attended Sessions (Conditional)	22.43	37.43
Average Attended Sessions (Unconditional)	0.77	24.25
% Students Missing Dosage Data	0.00	0.00

Note: The analysis sample is the sample of randomized students in Chicago (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

Our measure of student learning in this analysis is an index constructed from scores on end-of-year reading and math assessments. These assessments include the Renaissance STAR, i-Ready, Illinois Assessment of Readiness (IAR), PSAT, and SAT. Assessment availability varies by grade and school. For each student, we calculated a simple average of their available test scores, with each assessment score standardized using control-group scores within grade, subject, and school year, to create their index value. As shown below, end-of-year standardized test scores in reading and math—our primary outcomes of interest—are missing for approximately 14% of the HDT sample and 7% of the SHDT sample, meaning outcome data are available for the vast majority of students. Importantly, *t*-tests indicate no statistically significant differences in the rate of missingness between the treatment and control groups in any of the two samples.

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (linked in Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session. These estimates are presented in the table below.

In Chicago, for the 2023-24 school-year, we do not detect any statistically significant effects of participation (or being offered a chance to participate) in either the HDT or SHDT tutoring models on student test scores. The point estimates for HDT are approximately zero (both the ITT and TOT). For SHDT, the ITT point estimate, which gauges the impact of offering a student SHDT, is 0.07 SD and the TOT estimate, gauging the impact of receiving at least one session, is 0.10 SD but both are very imprecisely measured and are not statistically significant at the conventional levels.

Table 8: Differential attrition, Chicago 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	13.59%	13.54%	13.65%	.14
SHDT	7.16%	7.55%	6.62%	.48

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. The following assessments are included in the index outcome for Chicago: STAR, i-Ready, Illinois Assessment of Readiness (IAR), PSAT, and SAT. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p*-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request.*

The lack of statistically significant impacts from tutoring may be attributable to several factors:

1. Both HDT and SHDT students received substantially fewer sessions than intended—approximately 40 sessions (20 hours) on average, versus the target of 100 or more sessions.

2. In the HDT sample, 11% of control group students received tutoring at a dosage that was comparable to that of treatment group students, diluting estimated ITT effects.

3. School coordinators reported a high degree of access to personalized learning supports outside of tutoring, which could have weakened the contrast between treatment and control conditions. For example, over 74% of coordinators reported that the vast majority of their students had access to computer assisted learning platforms for literacy and/or math.

WHAT'S NEXT

For Chicago, the research team is currently analyzing the implementation and cost study data for the 2023-24 school year, as well as conducting additional impact analyses on secondary outcomes of interest. The 2023-24 Chicago data will be pooled with data from other sites and years for the forthcoming personalized treatment effect analysis.

Chicago Public Schools has continued partnering with PLI during the 2024-25 school year. The research team will report findings for the current school year once data collection and analysis are complete.

Table 9: Impact estimates, Chicago 2023-24

	HDT		SHDT	
	ITT	TOT	ITT	TOT
Estimate	0.005	0.008	0.067	0.102
Std. Error	0.034	0.051	0.051	0.077
***p*-value**	0.881	0.881	0.183	0.185
95% CI lower bound	-0.062	-0.092	-0.032	-0.049
95% CI upper bound	0.072	0.107	0.167	0.253
N	2,181	2,181	960	960
Control Mean	-0.021	-0.021	0.006	0.006
Treatment Mean	-0.071	-0.071	-0.119	-0.119
R^2	0.503	0.503	0.501	0.500
Adj. R^2	0.458	0.458	0.473	0.471

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. The following assessments are included in the index outcome for Chicago: STAR, i-Ready, Illinois Assessment of Readiness (IAR), PSAT, and SAT. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

Preliminary findings from Fulton tutoring

THE PARTNERSHIP

Since 2022 the Personalized Learning Initiative has partnered with Fulton County Schools to implement and study the impacts of tutoring on learning. In 2023-2024, Fulton County Schools rolled out both math and reading tutoring within the district. 13 schools participated in the study: nine elementary schools, three middle schools, and one high school. We evaluated two types of tutoring implemented at those schools.

THE INTERVENTIONS

In Fulton, tutoring was provided either by tutors hired by one of eight contracted vendors, or by tutors who were hired as paraprofessionals and staffed at an elementary schools.

Students could receive one of two different tutoring intervention types. The difference between the two tutoring models was the ratio of students to tutors, and whether the model included edtech.

- **HDT - 4:1 tutoring** – Tutors worked with students in groups of four (or fewer)
- **SHDT - 8:1 tutoring with edtech** – Tutors worked with students in groups of eight (or fewer), with an integrated edtech platform for additional personalized instruction and practice opportunities

Both models were offered at all study schools. Both models provided tutoring during the school day in classrooms with an in-person tutor. Both models offered at least 120 minutes of tutoring per week over 19–32 weeks. In elementary schools, both models were used to teach either literacy or math in a school; all middle and high schools focused on math. Two vendors—Saga Education and Lightning Squad—provided technical assistance and curriculum for the math and literacy tutoring curriculums, respectively.

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** - In all 13 schools, due to limited tutoring slots, not all eligible students could be provided with tutoring. We randomly assigned eligible students to one of three conditions—HDT, SHDT, or a business as usual group—using individual-level, classroom-level, and teacher-level lotteries. Because of this randomization, for a large enough sample, any differences in end-of-year test scores between the different groups of students are attributable directly to the interventions.
- **Surveys with tutors** - The PLI team obtained a roster of all tutors in PLI study schools in Fulton County in the fall of 2023 (N = 65 tutors) and again in the spring of 2024 (N = 71 tutors). In both waves of the survey, 82% of tutors responded.
- **Surveys and interviews with coordinators** - A school staff member responsible for coordinating the tutoring program in each of the PLI study schools was surveyed in the spring of 2024 (N = 13 school staff) and 92% responded. In addition, we also conducted interviews with a subsample of 10 coordinators. These interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation.

THE STUDENTS

At some schools, all students in non-accelerated classes in the grades selected for tutoring were eligible for the intervention. In other schools, only students one or more years behind in the grades selected for tutoring were eligible for the intervention.

The 2023-24 randomized sample includes just over 3,500 students. The sample is 97% students of color (mostly Black). As shown in the below table, the randomized sample is well-balanced on observable characteristics for the HDT sample. The SHDT sample is balanced on all the displayed dimensions, however overall balance is rejected due to imbalances in the share of students attending grades 7 and 8 (not shown in the table for space).

Table 10: Baseline balance, Fulton 2023-24, HDT

Covariate	Control Mean N = 1202	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	9.83	0.09	0.441	0.474	1849
% Male	0.49	-0.03	0.372	0.378	1851
% White	0.17	-0.02	0.164	0.174	1851
% Black	0.83	0.01	0.412	0.409	1851
% Hispanic	0.2	-0.03	0.112	0.129	1851
% Native American	0.03	0	0.985	0.987	1851
% Other Race	0.01	0	0.97	0.948	1851
% English as a Second Language	0.1	-0.01	0.681	0.694	1851
% Receiving Free/Reduced Lunch	1	0	0.836	0.516	1851
% Homeless	0.03	0	0.812	0.82	1851
Number of Days Attended	151.08	2.92	0.145	0.14	1656
Overall GPA	3.12	-0.05	0.185	0.229	1057
Overall GPA x Math	2.54	-0.04	0.221	0.26	1057
Overall GPA x ELA	0.58	-0.01	0.599	0.606	1057
Latest Available Math Score	0.01	-0.04	0.614	0.599	1831
Second Latest Available Math Score	0.02	-0.05	0.458	0.471	1671
Latest Available Reading Score	0.03	-0.01	0.851	0.866	1820
Second Latest Available Reading Score	0.03	0.01	0.945	0.949	1637
% Grade 6	0.15	-0.03	0.671	0.689	1851
% Grade 7	0.21	-0.02	0.706	0.755	1851
% Grade 8	0.05	0.05	0.423	0.578	1851
***F*-Test - Baseline Cov.**			0.819	0.993	1851

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in Fulton who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and control groups. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test. For brevity, we only show grade-level indicators that vary within randomization blocks.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly reassign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g., HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: **** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 11: Baseline balance, Fulton 2023-24, SHDT

Covariate	Control Mean N = 1096	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	10.15	0.08	0.412	0.448	1961
% Male	0.49	0.02	0.644	0.634	1963
% White	0.18	0.01	0.801	0.933	1963
% Black	0.81	-0.02	0.64	0.905	1963
% Hispanic	0.21	0.02	0.645	0.905	1963
% Native American	0.03	0.01	0.265	0.321	1963
% Other Race	0.01	0	0.567	0.558	1963
% English as a Second Language	0.1	0.04	0.471	0.846	1963
% Receiving Free/Reduced Lunch	1	0	0.424	0.409	1963
% Homeless	0.03	0.01	0.332	0.352	1963
Number of Days Attended	151.11	-1.67	0.39	0.386	1769
Overall GPA	3.12	-0.04	0.437	0.471	1241
Overall GPA x Math	2.53	-0.03	0.507	0.544	1241
Overall GPA x ELA	0.58	-0.01	0.42	0.389	1241
Latest Available Math Score	-0.02	0.11	0.256	0.312	1932
Second Latest Available Math Score	0.01	-0.05	0.618	0.665	1776
Latest Available Reading Score	0	0.02	0.839	0.871	1915
Second Latest Available Reading Score	0.02	0.05	0.673	0.719	1745
% Grade 6	0.16	0.05	0.243	0.263	1963
% Grade 7	0.23	-0.2	0.009***	0.009***	1963
% Grade 8	0.05	0.15	0.043**	0.086*	1963
***F*-Test - Baseline Cov.**			0***	0.016**	1963

Note: The analysis sample is composed of SHDT and BAU students randomized in a block in 2023-24 in Fulton who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with SHDT students are included. Reported p*-values test the difference in means for the SHDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test. For brevity, we only show grade-level indicators that vary within randomization blocks.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly reassign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: *** p < 0.01, ** p < 0.05, * p < 0.10.*

TAKE-UP AND DOSAGE

About three-quarters of students in both groups attended at least one session (75% for 4:1 HDT tutoring, 73% for 8:1 tutoring with edtech, SHDT). However, the recorded dosage was below expectations for both HDT and SHDT. Students participating in HDT received an average of 20.3 sessions, or 951 minutes, of tutoring per student over the year, and participating SHDT students received an average of 17.5 sessions, or 995 minutes, of tutoring over the year. The average conditional received dosage of about 16 hours per student for the year represents about one fifth to one third of the expected tutoring dosage that should have been received (expected dosage would have between 42 and 72 hours, depending on the model).

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

The outcome of interest is an index constructed from scores on end-of-year reading and math assessments—GMAS for grades 3–9, NWEA MAP for 9th grade, and iReady for grades K–8. Each assessment's scores are standardized using control-group scores in the same grade for that assessment. For each student, we calculated a simple average of their available test scores. Assessment availability varies by grade and school. The end-of-year standardized test scores in math and reading (depending on the tutored subject) are available for 83% of students, in HDT and SHDT as shown below. Importantly, *t*-tests indicate no statistically significant differences in the rate of missingness between the treatment and control groups in any of the two samples.

Table 12: Take-up and dosage, Fulton 2023-24

HDT Analysis Sample		
	BAU	HDT
N	1,202	649
% Received Treatment	1.58	75.35
Average Attended Sessions (Conditional)	8.32	20.26
Average Attended Sessions (Unconditional)	0.13	15.27
Average Scheduled Minutes (Conditional)	249.47	950.88
Average Scheduled Minutes (Unconditional)	3.94	716.46
% Students Missing Dosage Data	0.00	0.00

SHDT Analysis Sample		
	BAU	SHDT
N	1,096	867
% Received Treatment	0	72.90
Average Attended Sessions (Conditional)	0	17.54
Average Attended Sessions (Unconditional)	0	12.78
Average Scheduled Minutes (Conditional)	0	994.7
Average Scheduled Minutes (Unconditional)	0	725.09
% Students Missing Dosage Data	0	0.00

Note: The analysis sample is the sample of randomized students in Fulton (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

Table 13: Differential attrition, Fulton 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	17%	16.69%	17.16%	0.73
SHDT	16.86%	16.39%	17.22%	0.83

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p*-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request. The following assessments are included in the index outcome for Fulton: NWEA MAP, Georgia Milestones, i-Ready.*

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (see Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session. These estimates are presented in the table below.

In Fulton, for the 2023-24 school year, we do not find any statistically significant effects of participation (or being offered a chance to participate) in HDT on student test scores. However, our study found that SHDT significantly increased student test scores (*p*-value < 0.1 for both ITT and TOT): those who received SHDT learned 0.11 SD more over the year than those who did not receive tutoring but had access to all other status quo services, approximately 20% of the expected growth in a school year. We find that these SHDT impacts are driven by the students in grades 6–9.[13]

Across both tutoring models, we find large differences in impacts between elementary schools vs. middle schools and high schools. We consistently find that point estimates for the effect in grades K–5 are smaller than for grades 6–9.

In particular, the non-statistically significant point estimates we find at the high school level correspond to 42% additional learning (relative to not receiving tutoring) for HDT and SHDT. During site visits, it was informally reported that more frequent changes in schedules may have negatively affected the implementation of the program at the elementary level.

WHAT'S NEXT

The research team is still analyzing the implementation and cost study data for 2023-24, as well as refining the analysis of impacts and conducting additional impact analyses on secondary outcomes of interest. Additionally, the Fulton data will be pooled with other sites for the forthcoming personalized treatment effect analysis.

Fulton County has continued partnering with PLI during school year 2024-25. The research team will report findings for the current school year once data collection and analysis are complete.

[12] Based on the national-level average learning in ELA and math in for K-5 students, and math learning for 6-9th graders (see here, Bloom et al 2008).

[13] All estimates are based on early analysis and subject to future changes.

Table 14: Impact estimates, Fulton 2023-24

	HDT		SHDT	
	ITT	TOT	ITT	TOT
Estimate	0.009	0.012	0.078	0.11
Std. Error	0.033	0.044	0.043	0.06
***p*-value**	0.79	0.789	0.071	0.068
95% CI lower bound	-0.056	-0.074	-0.007	-0.008
95% CI upper bound	0.073	0.097	0.163	0.229
N	1851	1851	1963	1963
Control Mean	-0.034	-0.034	-0.056	-0.056
Treatment Mean	-0.016	-0.016	0.038	0.038
R^2	0.587	0.587	0.533	0.532
Adj. R^2	0.568	0.568	0.514	0.513

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. For blocks with clustered randomization, standard errors are clustered at the stratification level if there are fewer than 10 randomization units in a block, otherwise, standard errors are clustered at the randomization unit level. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation. The following assessments are included in the index outcome for Fulton: NWEA MAP, Georgia Milestones, i-Ready.

Preliminary findings from Greenville tutoring

THE PARTNERSHIP

In the 2023-24 school year, the Personalized Learning Initiative research team partnered with Greenville County Schools, as well as a virtual tutoring provider (Littera) and an edtech platform provider. We piloted two types of personalized learning interventions in three middle schools, and we conducted a study to explore impacts.

THE INTERVENTIONS

In Greenville, we piloted two personalized learning interventions—i) virtual tutoring and ii) pure edtech tutoring—both designed to provide mathematics support to middle schoolers.

- **HDT - 3:1 virtual tutoring** - Virtual tutors worked with groups of three students via an online platform and used the edtech product to provide instructional materials based on their students' unique needs.
- **SHDT - pure edtech** - Greenville's more sustainable, lower-cost model was having the "tutored" students work independently on a personalized edtech product. The edtech product offered individualized content and practice opportunities for students based on their unique needs.

Both interventions were to be delivered during students' homeroom period for 30 minutes, three times a week, over the course of 20 weeks.

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** - In all three schools, all middle-grade students were eligible to receive either virtual tutoring or edtech. Resources were available to serve some, but not all, middle school students eligible for math tutoring, allowing us to conduct a study in which students were assigned to receive tutoring using a fair lottery system. In each homeroom, we randomly assigned students to one of three conditions—virtual tutoring, edtech, or a business as usual group—using student-level randomization. Because of this fair lottery, any differences in end of year test scores between the control and treatment groups were attributable directly to the interventions.
- **Surveys with tutors** - The PLI team obtained a roster of all tutors in PLI study schools in Greenville in the the Fall of 2023 (N = 171) and the Spring of 2024 (N = 186 tutors). Response rates were high with 86% of tutors responding in Fall 2023 and 94% responding in the Spring of 2024.
- **Surveys and interview with coordinators** - A school staff member responsible for coordinating the tutoring program in each of the PLI study schools was surveyed and invited to complete an interview in the Spring of 2024. The interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation and resource-use.

THE STUDENTS

All middle school students at the three schools were eligible for the two interventions. Our analysis sample includes 1,878 students for whom we observe end-of-year test scores in the tutoring subject area (623 in HDT, 629 in SHDT, and 626 in BAU[14]) with students evenly distributed between 6th, 7th, and 8th grade. Of the participants, more than 85% were students of color, and more than 40% were English Language Learners. As shown in the below tables, the HDT analysis sample is well-balanced both overall and for each baseline covariate. For the SHDT sample, we observe imbalances in several variables, including grade and race categories. The *p*-value from the traditional joint *F*-test suggests overall imbalance at the 10% significance level. However, the randomization inference *p*-value suggests that the observed level of imbalance is not statistically significantly greater than what would be expected given the classroom-level random assignment conducted.

[14] Note that the 626 students in BAU are the same in both the HDT and SHDT tables below.

Table 15: Baseline balance, Greenville 2023-24, HDT

Covariate	Control Mean N = 626	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.27	0	0.995	0.995	1249
% Male	0.53	0	0.902	0.903	1236
% Black	0.29	-0.01	0.797	0.798	1236
% Hispanic	0.58	-0.01	0.759	0.766	1236
% White	0.09	0.02	0.207	0.219	1236
% Multi	0.04	-0.01	0.366	0.384	1236
% Other	0.01	0	0.527	0.618	1236
% English as a Second Language	0.42	0.03	0.314	0.328	1236
% Diverse Learner	0.19	0.02	0.432	0.44	1236
Latest Baseline Math Score	0.01	0.08	0.158	0.157	1109
6th Grade	0.34	0.01	0.112	0.142	1249
7th Grade	0.32	0	0.794	0.795	1249
8th Grade	0.34	-0.01	0.165	0.167	1249
***F*-Test - Baseline Cov.**			0.337	0.441	1249

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in Greenville who have at least one end-of-year test score in their tutored subject. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest available score is defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and control groups. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g. HDT and BAU to test a hypothesis for HDT vs BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: **** p *< 0.01, *** p *< 0.05, ** p *< 0.10.*

Table 16: Baseline balance, Greenville 2023-24, SHDT

Covariate	Control Mean N = 626	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.27	-0.05	0.085*	0.091*	1255
% Male	0.53	-0.01	0.794	0.813	1243
% Black	0.29	-0.06	0.019**	0.028**	1243
% Hispanic	0.58	0.01	0.654	0.642	1243
% White	0.09	0.04	0.049**	0.048**	1243
% Multi	0.04	0.02	0.215	0.227	1243
% Other	0.01	0	0.536	0.544	1243
% English as a Second Language	0.42	0.02	0.484	0.483	1243
% Diverse Learner	0.19	0.01	0.567	0.566	1243
Latest Baseline Math Score	0.01	0.03	0.577	0.572	1104
6th Grade	0.34	0.01	0.062*	0.084*	1255
7th Grade	0.32	0	0.805	0.824	1255
8th Grade	0.34	-0.01	0.225	0.26	1255
***F*-Test - Baseline Cov.**			0.085*	0.151	1255

Note: The analysis sample is composed of SHDT and BAU students randomized in a block in 2023-24 in Greenville who have at least one end-of-year test score in their tutored subject. Reported p*-values test the difference in means for the SHDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest available score is defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: **** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

TAKE-UP AND DOSAGE

As shown below, in the HDT sample, approximately 92% of students assigned to the treatment group received the treatment, defined as participating in at least one tutoring session. None of the HDT control students received tutoring. In our SHDT analysis sample, about 89% of treatment students and about 12% of control students ended up receiving tutoring.

For students assigned to treatment, conditional on receiving at least one session, students received about 20 sessions (20.19 for HDT, 19.85 for SHDT). Note that the target dosage was 60 sessions for both treatment groups.

Table 17: Take-up and dosage, Greenville 2023-24

HDT Analysis Sample		
	BAU	HDT
N	626	623
% Received Treatment	0	92.3
Average Attended Sessions (Conditional)	0	20.19
Average Attended Sessions (Unconditional)	0	18.64
Average Scheduled Minutes (Conditional)	0	605.84
Average Scheduled Minutes (Unconditional)	0	559.17
% Students Missing Dosage Data	0	0

SHDT Analysis Sample		
	BAU	SHDT
N	626	629
% Received Treatment	12.14	89.19
Average Attended Sessions (Conditional)	3.54	19.85
Average Attended Sessions (Unconditional)	0.43	17.71
Average Scheduled Minutes (Conditional)	106.18	595.56
Average Scheduled Minutes (Unconditional)	12.89	531.18
% Students Missing Dosage Data	0	0

Note: The analysis sample is the sample of randomized students in Greenville (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

Table 18: Differential attrition, Greenville 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	6.44%	6.32%	6.57%	0.84
SHDT	6.62%	6.68%	6.57%	0.71

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p*-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request.*

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

Our measure of student learning in this analysis is an index constructed from student scores on two end-of-year math assessments: the South Carolina College- and Career-Ready Assessments (SC-READY) and the Mastery View Predictive Assessments (MVPA). We created this index by calculating the simple average of the two test scores, each standardized within grade using the control group's mean and standard deviation. As shown below, the constructed index based on end-of-year standardized math test scores—the primary outcome—is missing for approximately 6% of students, indicating that outcome data are available for the vast majority of the sample. Importantly, *t*-tests indicate no statistically significant differences in the rate of missingness between the treatment and control groups in any of the two samples.

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (see Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session. The table below presents the ITT and TOT impact estimates for both the HDT and SHDT in Greenville for the 2023–24 school year.

In Greenville, we saw statistically significant and positive impacts of virtual 3:1 tutoring (HDT) on student learning as measured by an index of end-of-year math scores. On average, students who were offered an opportunity to participate in virtual tutoring gained an additional 0.10 standard deviations in the end of year math score index compared to those randomized into the control group and the treatment effect for students who ended up participating in virtual tutoring is estimated to be 0.11 standard deviations (*p*-value = 0.006).[15] This effect is equivalent to approximately 32% of the math an average middle school student learns in a year.[16]

In Greenville, we were not able to detect any statistically significant impact from engaging with edtech (SDHT) on student math learning as measured by the end of year math scores. Additionally, according to observations from the technical assistance teams and interviews with school coordinators, students in the control group in some schools may have engaged with other edtech products during homeroom time (in addition to the 12% of BAU students who received the SHDT intervention), which could have weakened the contrast between the two groups.

[15] These estimates are based on early analysis and subject to future changes.

[16] An average middle-school student learns about 0.34 standard deviations in math in nationally normed tests (see here, Bloom et al 2008).

Table 19: Impact estimates, Greenville 2023-24

	HDT		SHDT	
	ITT	TOT	ITT	TOT
Estimate	0.102	0.11	0.026	0.034
Std. Error	0.037	0.04	0.038	0.049
***p*-value**	0.006	0.006	0.493	0.492
95% CI lower bound	0.029	0.024	-0.048	-0.07
95% CI upper bound	0.175	0.196	0.1	0.137
N	1249	1249	1255	1255
Control Mean	0.006	0.006	0.006	0.006
Treatment Mean	0.133	0.133	0.062	0.062
R^2	0.613	0.614	0.57	0.57
Adj. R^2	0.569	0.569	0.521	0.521

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

WHAT'S NEXT

2023-24 was the only year of PLI participation for Greenville, so no new data is forthcoming for this site. However, the research team is still analyzing the implementation and cost study data for 2023-24, as well as conducting additional impact analysis on secondary outcomes. In addition, the Greenville data will be pooled with other sites for the forthcoming personalized treatment effect analysis.

Preliminary findings from Guilford tutoring

THE PARTNERSHIP

In the 2023-24 school year, the PLI research team partnered with Guilford County Schools in North Carolina, as well as two tutoring providers: University of North Carolina Greensboro (UNCG), through its Institute for Partnerships in Education (IPiE), along with Kelly Services. This partnership was facilitated by The Innovation Project in North Carolina. We piloted two types of personalized learning interventions in two middle schools, and we also conducted a study to explore impacts. Tutoring began in February 2024 and lasted 11 weeks.

THE INTERVENTIONS

In Guilford, we piloted two personalized learning interventions—one more resource intensive, the other less so—both were designed to provide mathematics support to middle schoolers. Tutors for both interventions interacted with students in person and used a combination of teacher- and district-developed materials for personalization to the unique needs of each student. The tutors pushed into a classroom and worked with students who were grouped by the school's tutoring coordinator. Both interventions were designed to be delivered during students' class time for 30 minutes, two to three times a week (every other day), over the course of 10–11 weeks for an intended dosage of 750–825 minutes in total.

The key differences between the two interventions offered in each school were the student-to-tutor ratios and the use of edtech:

- **HDT - 4:1 tutoring** - Tutors supported four HDT students at a time.
- **SHDT - 8:1 tutoring with edtech** - Tutors supported 8 students at a time, while leveraging the edtech platform Dreambox.

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** - In both participating schools, most middle school students were eligible to receive either HDT or SHDT.[17] However, due to limited funding, not all students were able to be assigned to these interventions during the school day. We randomly assigned students to one of three conditions—HDT, SHDT, or a business as usual (BAU) group—using classroom-level randomization. Because of this fair lottery, for a large enough sample, any differences in end of year test scores between the control and treatment groups were attributable directly to the interventions in expectation.
- **Surveys with tutors** - The PLI team obtained a roster of all tutors in PLI study schools in Guilford in the spring of 2024 (N = 16 tutors). Response rates were high, with 88% of tutors responding.
- **Surveys and interview with coordinators** - A school staff member responsible for coordinating the tutoring program in each of the PLI study schools was surveyed and interviewed in the spring of 2024. The interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation and resource-use.

THE STUDENTS

As mentioned above, the majority of middle school students at the two schools were eligible for the two interventions. The randomized group included 1,415 students, with slightly more students in 6th and 8th grade. Of the participants, approximately 84% were students of color, and 15% were English Language Learners.

[17] Typically, the few classes excluded from randomization were excluded due to being resource or other special needs classes where the intervention would be deemed disruptive.

Despite the randomized assignment, the combination of classroom-level random assignment and the smaller effective sample size (56 classrooms) resulted in an imbalanced sample. In particular, by chance, treated students were more likely at baseline to have higher prior year GPAs, higher prior year test scores (both in benchmark assessments and in state standardized assessments), and were likelier to be enrolled in advanced math classes. While we statistically control for the baseline differences, this imbalance implies that we can be less confident that the difference in outcomes between the treatment and control students was driven purely by tutoring.

In particular, if unobserved residual differences are imbalanced in the same direction and positively correlated with the follow up test score outcomes, the impact estimate will be biased upward. In other words, if students who were higher performing at baseline tend to have unobservable characteristics, such as more at-home supports, that are also positively correlated with future test scores but are not fully adjusted for by our set of baseline control variables, some of the treatment-control difference in outcomes could be due to these unobservable and uncontrolled for factors.

TAKE-UP AND DOSAGE

In the HDT sample, approximately 98% of students assigned to the treatment group received the treatment, defined as participating in at least one tutoring session. Only 2% of HDT control students received tutoring. In our SHDT analysis sample, about 93% of treatment students and about 1% of control students ended up receiving tutoring. Students in both the HDT and SHDT groups each received, on average, about 10-12 sessions of the intervention, over the course of the semester, conditional on attending at least one session.

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

Our measure of student learning in this analysis is an index constructed from student scores on two end-of-year math assessments: a standardized state assessment, either the North Carolina End of Grade (EOG) or End of Course (EOC) assessment, depending on the student's mathematics grade level, and their NWEA MAP math score. We created this index by calculating the simple average of the two test scores, standardized within grade, subject, school-year, and assessment. As shown below, this primary outcome index is available for 98% of students.

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (see Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session.

In Guilford, we did not find statistically significant impacts of HDT on student learning as measured by an index of end-of-year math scores. On average, students who attended at least one session of HDT gained an additional 0.039 standard deviations in end-of-year math scores compared to their peers who did not receive tutoring of either kind. However, we cannot reject the null hypothesis that the true effect is zero at a 5 percent level of significance. On the other hand, SHDT students who received at least one session had a significant 0.15 SD difference in outcomes, relative to the control group. As mentioned above, we cannot be certain that these estimates capture only the causal impact of tutoring, since the treatment students were already outperforming the control students at baseline by a large margin.

WHAT'S NEXT

2023-24 was the only year of PLI participation for Guilford. However, the research team is still analyzing the implementation and cost study data for 2023-24, as well as conducting additional impact analysis on secondary outcomes. In addition, the Guilford data will be pooled with other sites for the forthcoming personalized treatment effect analysis.

Table 20: Baseline balance, Guilford 2023-24, HDT

Covariate	Control Mean N = 631	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.14	-0.15	0.541	0.588	1018
% Male	0.57	-0.06	0.039**	0.082*	1018
% White	0.15	-0.02	0.324	0.422	1018
% Black	0.5	-0.05	0.179	0.255	1018
% Asian	0.07	0.04	0.061*	0.125	1018
% Other	0.07	-0.01	0.678	0.71	1018
% Hispanic	0.22	0.04	0.081*	0.136	1018
% English as a Second Language	0.17	-0.01	0.663	0.709	1018
% Disability	0.14	-0.05	0.138	0.199	1018
Number of Days Attended	155.03	2.22	0.138	0.224	934
Overall GPA	2.56	0.28	0.041**	0.102	930
Latest Available Math Score	0	0.32	0.03**	0.077*	1017
Second Latest Available Math Score	-0.01	0.26	0.097*	0.183	989
Latest Available Reading Score	0	0.29	0.018**	0.052*	1014
Second Latest Available Reading Score	0	0.26	0.035**	0.074*	987
Grade 6	0.37	0	0.986	0.987	1018
Grade 7	0.21	0.12	0.404	0.496	1018
Grade 8	0.42	-0.12	0.373	0.446	1018
***F*-Test - Baseline Cov.**			0***	0.506	1018

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in Guilford who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and control groups. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly reassign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g., HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: **** $p < 0.01$*, *** $p < 0.05$*, ** $p < 0.10$*.*

Table 21: Baseline balance, Guilford 2023-24, SHDT

Covariate	Control Mean N = 631	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.14	-0.04	0.868	0.873	999
% Male	0.57	-0.06	0.027**	0.07*	999
% White	0.15	0.04	0.128	0.192	999
% Black	0.5	-0.13	0***	0.003***	999
% Asian	0.07	0.03	0.142	0.218	999
% Other	0.07	-0.01	0.511	0.533	999
% Hispanic	0.22	0.07	0.039**	0.087*	999
% English as a Second Language	0.17	-0.03	0.343	0.441	999
% Disability	0.14	-0.05	0.142	0.242	999
Number of Days Attended	155.03	4.46	0.004***	0.012**	914
Overall GPA	2.56	0.48	0***	0.002***	911
Latest Available Math Score	0	0.66	0***	0.003***	999
Second Latest Available Math Score	-0.01	0.56	0***	0.006***	974
Latest Available Reading Score	0	0.52	0***	0.003***	996
Second Latest Available Reading Score	0	0.46	0***	0.003***	972
Grade 6	0.37	-0.11	0.415	0.491	999
Grade 7	0.21	0.22	0.092*	0.175	999
Grade 8	0.42	-0.11	0.363	0.388	999
***F*-Test - Baseline Cov.**			0***	0.04**	999

Note: The analysis sample is composed of SHDT and BAU students randomized in a block in 2023-24 in Guilford who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with SHDT students are included. Reported p*-values test the difference in means for the SHDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows:*
*** p < 0.01, ** p < 0.05, * p < 0.10.*

Table 22: Take-up and dosage, Guilford 2023-24

HDT Analysis Sample		
	BAU	HDT
N	631	387
% Received Treatment	2.22	98.19
Average Attended Sessions (Conditional)	6.86	11.74
Average Attended Sessions (Unconditional)	0.27	11.53
Average Scheduled Minutes (Conditional)	205.71	352.34
Average Scheduled Minutes (Unconditional)	8.03	345.97
% Students Missing Dosage Data	0	0

SHDT Analysis Sample		
	BAU	SHDT
N	631	368
% Received Treatment	1.11	93.48
Average Attended Sessions (Conditional)	4	10.29
Average Attended Sessions (Unconditional)	0.14	9.62
Average Scheduled Minutes (Conditional)	120	308.81
Average Scheduled Minutes (Unconditional)	4.14	288.67
% Students Missing Dosage Data	0	0

Note: The analysis sample is the sample of randomized students in Guilford (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

Table 23: Differential attrition, Guilford 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	2%	2%	2%	.32
SHDT	2%	2%	2%	.52

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. Tests included in the index score are NWEA MAP test and North Carolina EOG/EOC state assessment. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request.

Table 24: Impact estimates, Guilford 2023-24

	HDT		SHDT	
	ITT	TOT	ITT	TOT
Estimate	0.037	0.039	0.137	0.151
Std. Error	0.048	0.05	0.045	0.049
***p*-value**	0.439	0.444	0.002	0.004
95% CI lower bound	-0.057	-0.062	0.050	0.053
95% CI upper bound	0.131	0.140	0.225	0.25
N	1018	1018	999	999
Control Mean	0.002	0.002	0.002	0.002
Treatment Mean	0.24	0.24	0.64	0.64
R^2	0.756	0.756	0.795	0.794
Adj. R^2	0.748	0.748	0.788	0.787

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. Tests included in the index score are NWEA MAP test and North Carolina EOG/EOC state assessment. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

Preliminary findings from Miami tutoring

THE PARTNERSHIP

In the 2023-24 school year, the PLI research team partnered with Miami-Dade County Public Schools. We piloted HDT in nine middle schools and SHDT in one school, starting in the second semester (January-May 2024). We conducted a study to explore the impacts of HDT from this half-year pilot program. With only four classrooms that offered SHDT, we decided not to estimate SHDT impacts.

THE INTERVENTION

The nine Miami schools in our analysis implemented:[18]

- **HDT - 4:1 tutoring** - Two math tutors pushed into either a core or remedial (foundational) math class. Students in these classrooms were divided into three groups based on performance on previous math assessments. During the class period, each group rotated through a teacher-led, tutor-led, and an edtech station, spending 20-30 minutes at each. Each tutor supported a maximum of about four students per group.

Tutoring was designed to be 30 minutes, meeting twice one week and three times the next week for an average of 2.5 times per week, for roughly 19 weeks.

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** - In the nine participating schools, all middle-grade students were eligible for tutoring services. However, due to budget constraints, not all classrooms could be assigned to receive tutoring. Classrooms were randomly assigned to one of two conditions—HDT or a business as usual (BAU) control group—using classroom-level randomization. Classrooms assigned to the HDT condition received two tutors each, while BAU classrooms did not receive any tutoring support. Because assignment was determined through a fair lottery, any observed differences in end-of-year test scores between students in HDT and BAU classrooms can be causally attributed to the tutoring intervention.

- **Surveys with tutors** - The PLI team obtained a roster of all tutors in PLI study schools in Miami in the spring of 2024 (N = 17 tutors). Response rates were high with 100% of tutors responding.

- **Surveys and interview with coordinators** - A school staff member responsible for coordinating the tutoring program in each of the PLI study schools was surveyed in the spring of 2024 (N = 9 school staff members) and 100% responded. In addition, we conducted interviews with a subsample of three of these school coordinators. These interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation and resource-use.

THE STUDENTS

The randomized group included about 2,400 students in 6th, 7th and 8th grade. Of the included participants, 97% were students of color (75% Hispanic), and 32% were English Language Learners.

As shown in the tables below, we observe some statistically significant differences between treatment and control groups on certain baseline characteristics, such as race. The *p*-value from the traditional joint *F*-test indicates overall imbalance, however, the randomization inference *p*-value suggests that the observed level of imbalance is not statistically significantly greater than what would be expected given the classroom-level random assignment conducted.

[18] The sustainable high dosage tutoring model was piloted at only one school in four classrooms. In this model, one tutor pushed into each classroom, increasing the number of students the tutor worked with to eight students. Due to limited sample size, the SHDT model is excluded from the 2023–24 impact analysis for Miami.

Table 25: Baseline balance, Miami 2023-24, HDT

Covariate	Control Mean N = 1214	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.21	0.03	0.821	0.862	2,134
% Male	0.52	0.01	0.75	0.806	2,134
% White	0.75	0.03	0.002***	0.014**	2,134
% Hispanic	0.75	0.03	0.018**	0.079*	2,134
% Other Race	0.26	-0.03	0.002***	0.013**	2,134
% English as a Second Language	0.33	0	0.925	0.938	2,134
% Receiving Free/Reduced Lunch	0.73	-0.01	0.727	0.796	1,900
% Diverse Learner	0.2	0	0.83	0.868	2,134
Overall GPA	2.65	-0.06	0.138	0.277	1,836
Latest Available Math Score	0.06	-0.02	0.782	0.81	2,088
Second Latest Available Math Score	-0.03	0.08	0.175	0.27	2,061
Latest Available Reading Score	-0.01	0.06	0.374	0.487	2,095
Second Latest Available Reading Score	0.01	0.06	0.244	0.348	2,065
% Grade 6	0.39	-0.07	0.383	0.523	2,134
% Grade 7	0.32	0.08	0.276	0.381	2,134
% Grade 8	0.29	-0.02	0.842	0.88	2,134
***F*-Test - Baseline Cov.**			0.001***	0.344	2,134

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in Miami who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and control groups. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly reassign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g., HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: ***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.*

TAKE-UP AND DOSAGE

Over 93% of students in classes assigned to tutoring attended at least one session. On average, participating students received 19 total tutoring sessions throughout the semester.

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

Our outcome measure is standardized math scores in the Florida Assessment of Student Thinking (FAST). As shown below, end of year math scores (the primary outcome of interest) are available for about 94% of the students in randomized classes. Importantly, *t*-tests indicate no statistically significant differences in the rate of missingness between the treatment and control groups.

Table 26: Take-up and dosage, Miami 2023-24

HDT Analysis Sample		
	BAU	**HDT**
N	1,214	920
% Received Treatment	1.98	93.15
Average Attended Sessions (Conditional)	17.83	19.36
Average Attended Sessions (Unconditional)	0.35	18.04
Average Scheduled Minutes (Conditional)	535	580.85
Average Scheduled Minutes (Unconditional)	10.58	541.08
% Students Missing Dosage Data	0.00	0.00

Note: The analysis sample is the sample of randomized students in Miami (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. Only HDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

Table 27: Differential attrition, Miami 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	6.03%	6.12%	5.96%	0.64

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p*-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request. The only assessment included in the index outcome for Miami is the Florida Assessment of Student Thinking (FAST).*

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (see Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session. These estimates are presented in the table below.

The estimated TOT and ITT effects from the half-year pilot program in Miami are positive but not statistically significant at conventional levels. Given the limited duration and scale of the pilot, these preliminary findings are nonetheless encouraging. For the upcoming school year, we are increasing program dosage by extending implementation to a full academic year. In addition, we are expanding the sample by doubling the number of participating schools, which is expected to increase statistical power and improve our ability to detect program impacts.

WHAT'S NEXT

PLI participation in Miami doubled to 18 schools in the 2024-25 school year. In collaboration with Miami-Dade County Public Schools, the research team is additionally in the planning phases for the 2025-26 school year. As new data becomes available, the research team will continue to analyze impact as well as cost and implementation data. Additionally the Miami data will be pooled with other sites for the forthcoming personalized treatment effect analysis.

Table 28: Impact estimates, Miami 2023-24

	HDT	
	ITT	**TOT**
Estimate	0.046	0.050
Std. Error	0.031	0.034
***p*-value**	0.137	0.141
95% CI lower bound	-0.015	-0.017
95% CI upper bound	0.106	0.116
N	2,134	2,134
Control Mean	0.000	0.000
Treatment Mean	0.086	0.086
R^2	0.489	0.490
Adj. R^2	0.473	0.474

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation. The following assessments are included in the index outcome for Miami: The only assessment included in the index outcome for Miami is the Florida Assessment of Student Thinking (FAST).

Preliminary findings from New Mexico tutoring

THE PARTNERSHIP

Since 2022, the New Mexico Public Education Department (NMPED) has been working with Saga Education and the PLI team to stand up virtually-delivered live tutoring for middle school math across the state and study its impacts on student learning. In school year 2023-2024, 19 middle schools participated in the study. Eight of these schools were in rural areas and five were in small towns.

THE INTERVENTION

In 2023-24, New Mexico provided one personalized learning intervention:

- **HDT - virtual 4:1 tutoring** - New Mexico rolled out virtual HDT at volunteering schools across the state. Tutoring took place in a classroom during the school day, with students connecting to their tutor through individual devices (laptops, desktop PCs, or tablets). A proctor, most often a math teacher, proctored these classrooms. Whenever possible, tutors worked with groups of 3 or 4 students at a time. Tutoring was scheduled for at least 90 minutes per week over 32-36 weeks.

NMPED hired their own virtual tutors. The program used the math curriculum developed by Saga

Education. Saga provided technical assistance support and trained tutors with NMPED.

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** - In all 19 schools, all middle-grade students were eligible to receive virtual tutoring. However, due to limited funding, not all students could be provided this intervention. We randomly assigned students to one of two conditions—virtual tutoring or a business-as-usual group—using grade-level randomization for the majority of the participating students, although at one school the principal requested individual-level randomization.

  In the 18 schools with grade-level random assignment, HDT was offered to all the students in the randomly selected grade level—6th, 7th, or 8th. Because we used a fair lottery for both grade level and individual level random assignment, the differences in end-of-year test scores between the tutored and non-tutored students across all 19 schools can be attributed directly to tutoring and not other factors, such as pre-existing achievement differences.

- **Surveys with tutors** - The PLI team obtained a roster of all tutors working with PLI study schools in New Mexico in the fall of 2023 (N = 62) and in the spring of 2024 (N = 58 tutors). Response rates were high with 90% of tutors responding in both waves of the survey.

- **Surveys and interview with coordinators** - A school staff member responsible for coordinating the tutoring program in each of the PLI study schools was surveyed in the spring of 2024 (N = 19 school staff members) and 90% responded. In addition, we also conducted interviews with a sub-sample of eight of these school coordinators. These interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation and resource-use.

THE STUDENTS

The 19 schools participating in the study could elect to have all grades 6-8th be eligible for tutoring or only two of the three grades. All middle school students enrolled in those grades were eligible for the intervention. The randomized group included just over 1,500 students, with students distributed between 6th, 7th, and 8th grade. One school, however, was later excluded from the analysis due to non-compliance with the lottery results, as outlined by the pre-registered analysis plan (see Appendix 1). Of the included participants, 69% were students of color, (with about 18% being Native Americans) and 17% were English Language Learners. As shown in the below table, the randomized sample is well-balanced overall and on most observable characteristics. We observe small but statistically significant imbalances (at the 5% level) for race, with the treatment group being slightly more white and less Native or Hispanic.

While overall balance is rejected by the standard *F*-test, the randomization inference-based *F*-test is not statistically significant suggesting that the observed level of imbalance would not be uncommon given the grade-level assignment.

Table 29: Baseline balance, New Mexico 2023-24, HDT

Covariate	Control Mean N = 735	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.16	0	0.999	1	1208
% Male	0.49	0.02	0.336	0.357	1207
% Hispanic	0.55	-0.05	0.089*	0.145	1204
% Black	0.02	0	0.936	0.938	1200
% White	0.76	-0.02	0.01***	0.022**	1200
% Asian	0.01	0.01	0.181	0.2	1200
% Native American	0.19	0.03	0.025**	0.031**	1200
% Pacific Islander	0.06	-0.01	0.238	0.27	1200
% English as a Second Language	0.16	0	0.799	0.82	1140
% Diverse Learner	0.23	-0.03	0.306	0.335	1139
% Receiving Free/Reduced Lunch	0.56	0.04	0.272	0.308	798
% Homeless	0.02	0	0.581	0.606	1087
Latest Available Math Score	0.04	0	0.997	0.998	1147
Second Latest Available Math Score	-0.05	-0.05	0.599	0.621	1083
Latest Available Reading Score	0.02	0	0.985	0.99	1147
Second Latest Available Reading Score	0.04	-0.06	0.744	0.772	1069
Grade 6	0.29	0.01	0.929	0.917	1208
Grade 7	0.38	-0.05	0.838	0.852	1208
Grade 8	0.33	0.03	0.876	0.876	1208
***F*-Test - Baseline Cov.**			0***	0.409	1208

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in New Mexico who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and control groups. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly reassign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g. HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: *** p < 0.01, ** p < 0.05, * p < 0.10.*

TAKE-UP AND DOSAGE

The virtual tutoring program had high student participation, with 92% of students assigned to tutoring attending at least one session. On average, these students received about 64 sessions per student over the year or 2,754 minutes, which is over 40 hours.

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

The outcome of interest is the end of year state assessment New Mexico Measures of Student Success and Achievement (NM-MSSA), which is standardized using control group scores in the same grade. Assessment availability varies by grade and school. The end-of-year standardized math test scores are available for 83% of students. Importantly, *t*-tests indicate no statistically significant differences in the rate of missingness between the treatment and control groups in any of the two samples.

Table 30: Take-up and dosage, New Mexico 2023-24

HDT Analysis Sample		
	BAU	**HDT**
N	735	473
% Received Treatment	0.41	92.18
Average Attended Sessions (Conditional)	69.67	63.98
Average Attended Sessions (Unconditional)	0.28	58.98
Average Scheduled Minutes (Conditional)	2775	2753.64
Average Scheduled Minutes (Unconditional)	11.33	2538.24
% Students Missing Dosage Data	0	0

Note: The analysis sample is the sample of randomized students in New Mexico (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

Table 31: Differential attrition, New Mexico 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	17.37%	17.31%	17.42%	0.3

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. The following assessments are included in the index outcome for New Mexico: statewide summative assessment—MSSA. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p*-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request.*

Table 32: Impact estimates, New Mexico 2023-24

	HDT	
	ITT	TOT
Estimate	0.119	0.13
Std. Error	0.056	0.061
***p*-value**	0.032	0.034
95% CI lower bound	0.01	0.010
95% CI upper bound	0.229	0.251
N	1208	1208
Control Mean	0	0
Treatment Mean	0.11	0.11
R^2	0.601	0.601
Adj. R^2	0.584	0.584

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. For blocks with 10 or more randomization units, standard errors are clustered at the randomization unit level. For blocks with fewer than 10 randomization units, standard errors are clustered at the stratification level. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. The following assessments are included in the index outcome for New Mexico: statewide summative assessment—MSSA. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (see Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session. These estimates are presented in the table below.

In New Mexico, we saw statistically significant and positive impacts of virtual tutoring on student test scores. Our study found that the students who received HDT learned 0.13 SD more over the year than those who did not receive tutoring but had access to all other status quo services. The impact translates into approximately 38% of the expected growth in middle school math during an entire school year.[19,20]

WHAT'S NEXT

The research team is still analyzing the implementation and cost study data for 2023-24, as well as refining the analysis of impacts and conducting additional impact analyses on secondary outcomes of interest. Additionally, data from New Mexico will be pooled with other sites for the forthcoming personalized treatment effect analysis.

NMPED has continued partnering with PLI during school year 2024-25. The research team will report findings for the current school year once data collection and analysis are complete.

[19] These estimates are based on early analysis and subject to future changes.

[20] An average middle-school student learns about 0.34 standard deviations in math in nationally normed tests (see here, Bloom et al 2008).

Preliminary findings from Rocketship tutoring

THE PARTNERSHIP

Starting in 2023-2024, Rocketship Public Schools (a charter school network) and the Bay Area Tutor Corps (BATA) piloted two types of reading tutoring in two elementary schools in San Jose. The PLI team partnered with them to conduct an impact evaluation of the tutoring programs.

THE INTERVENTIONS

Rocketship and BATA implemented two high-dosage, small-group tutoring models: a traditional high dosage tutoring (HDT) model and a sustainable high dosage tutoring (SHDT) model. Both provided tutoring during the school day in classrooms with in-person tutors. Tutors were hired and trained by BATA.

- **HDT - 4:1 tutoring** - Tutors met with students in groups of up to four students at a time
- **SHDT - 8:1 tutoring** - Tutors met with students in groups of up to eight students at a time

For both types of tutoring, 30-minute sessions were scheduled four times per week (120 minutes per week) over 18–20 weeks—an expected dosage of more than 30 hours. The two Rocketship schools already provide their students with a lot of personalized learning: all students have a learning lab where they work on edtech platforms in math and reading and/or work with the schools' "individualized learning specialists." The HDT and SHDT were on top of this relatively high base of tutoring.

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** - In these two schools, all students in grades K–5 were eligible to receive tutoring. However, due to limited funding, not all students could be provided with tutoring. We randomly assigned students to one of three conditions—HDT, SHDT or a business-as-usual group—using grade-level randomization.

  In each school, two grades were randomly selected to receive HDT, and two others were selected to receive SHDT. The two remaining grades received all the standard instruction and support offered by the schools.

- **Surveys with tutors** - The PLI team obtained a roster of all tutors in PLI study schools in Rocketship in the spring of 2024 (N = 21 tutors). Response rates were high with 91% of tutors responding.

- **Surveys and interview with coordinators** - In Rocketship, coordinator interviews and surveys with key personnel were conducted to cover both study schools. The interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation and resource-use.

THE STUDENTS

As mentioned above, all elementary school students at the two schools were eligible for the two interventions. The randomized group included over 800 students, with students distributed between K–5th grades. Of the students in the analysis sample, 93% were students of color (primarily Hispanic), and 67% were English Language Learners. As shown in the below table, the randomized sample has a large imbalance in age, due to the randomization design. As evidenced by the randomization inference *p*-value, this imbalance is not large relative to the distribution of possible randomization outcomes given the chosen design. However, with almost two years of difference between the treatment and control group, the imbalance is quite large in absolute value and significant when evaluated using standard inference. We also observe that the treatment group has a larger share of male students, fewer white students, and a smaller number of school days attended in the prior year. These imbalances are reflected in the rejection of the hypothesis of balance by the standard *F*-test.

Table 33: Baseline balance, Rocketship 2023-24, HDT

Covariate	Control Mean N = 290	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	7.23	-1.84	0***	0.391	415
% Male	0.43	0.11	0.033**	0.057*	415
% Hispanic	0.82	0.02	0.561	0.945	415
% Asian	0.12	-0.01	0.786	0.886	415
% White	0.03	-0.02	0.084*	0.272	415
% Other Race	0.03	0.01	0.664	0.611	415
% Receiving Free/Reduced Lunch	0.72	-0.02	0.647	0.894	415
% English as a Second Language	0.62	0.05	0.327	0.824	415
% Diverse Learner	0.09	0.01	0.701	0.686	415
Number of Days Attended	159.83	-7.93	0.023**	0.288	283
Latest Available Reading Score	0	-0.09	0.416	0.318	412
Second Latest Available Reading Score	0.01	-0.04	0.719	0.597	403
% Grade K	0.2	0.36	0***	0.55	415
% Grade 1	0.27	0.16	0.002***	0.78	415
% Grade 3	0.34	-0.33	0***	0.89	415
% Grade 5	0.18	-0.2	0***	0.897	415
***F*-Test - Baseline Cov.**			0***	0.945	415

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in Rocketship who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and control groups. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g., HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the student level are also reported and statistical significance is denoted as follows: *** p < 0.01, ** p < 0.05, * p < 0.10.*

Table 34: Baseline balance, Rocketship 2023-24, SHDT

Covariate	Control Mean N = 290	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	7.23	1.6	0***	0.437	407
% Male	0.43	0.08	0.143	0.108	407
% Hispanic	0.82	-0.06	0.157	0.419	407
% Asian	0.12	0.08	0.046**	0.648	407
% White	0.03	0	0.913	0.879	407
% Other Race	0.03	-0.01	0.327	0.49	407
% Receiving Free/Reduced Lunch	0.72	0.04	0.473	0.776	407
% English as a Second Language	0.62	-0.08	0.142	0.507	407
% Diverse Learner	0.09	0.08	0.035**	0.11	407
Number of Days Attended	159.83	1.56	0.385	0.49	318
Latest Available Math Score	0.01	-0.25	0.083*	0.001***	313
Latest Available Reading Score	0	-0.25	0.033**	0.407	406
Second Latest Available Reading Score	0.01	-0.06	0.575	0.541	402
% Grade K	0.2	-0.29	0***	0.553	407
% Grade 1	0.27	-0.2	0***	0.797	407
% Grade 3	0.34	0.31	0***	0.683	407
% Grade 5	0.18	0.19	0.002***	0.726	407
***F*-Test - Baseline Cov.**			0***	0.88	407

Note: The analysis sample is composed of SHDT and BAU students randomized in a block in 2023-24 in Rocketship who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with SHDT students are included. Reported p-values test the difference in means for the SHDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting F-statistic from this regression. Missingness indicators are included in the regression model but not in the F-test.

*We report both the p-value and randomization inference p-values, to avoid distributional assumptions. To calculate the randomization inference p-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and estimate the corresponding test-statistic (p-value for pairwise tests, p-value of the F-statistic for joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI p-value. p-values clustered at the student level are also reported and statistical significance is denoted as follows: *** p < 0.01, ** p < 0.05, * p < 0.10.*

TAKE-UP AND DOSAGE

All students assigned to tutoring attended at least one session and both models registered similar dosage. Students who participated in tutoring averaged a total of 39.6 sessions for HDT and 37 for SHDT, or 1,118 and 1,111 minutes, respectively. This is approximately 18.5 hours of tutoring.

Table 35: Take-up and dosage, Rocketship 2023-24

HDT Analysis Sample		
	BAU	HDT
N	290	125
% Received Treatment	0	100
Average Attended Sessions (Conditional)	0	39.61
Average Attended Sessions (Unconditional)	0	39.61
Average Scheduled Minutes (Conditional)	0	1188.24
Average Scheduled Minutes (Unconditional)	0	1188.24
% Students Missing Dosage Data	0	0

SHDT Analysis Sample		
	BAU	SHDT
N	290	117
% Received Treatment	0	100
Average Attended Sessions (Conditional)	0	37.04
Average Attended Sessions (Unconditional)	0	37.04
Average Scheduled Minutes (Conditional)	0	1111.28
Average Scheduled Minutes (Unconditional)	0	1111.28
% Students Missing Dosage Data	0	0

Note: The analysis sample is the sample of randomized students in Rocketship (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

Our measure of student learning in this analysis is an index constructed from student scores on two end-of-year reading assessments: Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and the Northwest Evaluation Association Measures of Academic Progress (NWEA MAP) for grades K–5. Each assessment is standardized using control group scores in the same grade. Assessment availability varies by grade and school. As shown below, the index of end-of-year standardized test scores in reading (the primary outcome of interest) is available for 69% of students in the HDT sample and for 70% of students in the SHDT sample. While test scores are available for approximately 95% of students, the missingness in the analysis sample is generated by the lack of control group students in particular grades, which are necessary for outcome standardization. This occurred if randomly both grades across the two schools were assigned to a tutoring treatment.

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (see Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session. These estimates are presented in the table below.

We find no significant impact of the additional HDT or SHDT, with small, positive estimates of 0.06 SD and 0.08 SD. These effect sizes are equivalent to approximately an 8% and 10% increase in learning for an average elementary school student in reading, respectively.

Table 36: Differential attrition, Rocketship 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	31.40%	58.61%	4.29%	0.42
SHDT	30.07%	58.06%	4.29%	0.32

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p*-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request.*

Table 37: Impact estimates, Rocketship 2023-24

	HDT		SHDT	
	ITT	TOT	ITT	TOT
Estimate	0.061	0.061	0.079	0.079
Std. Error	0.065	0.065	0.054	0.054
***p*-value**	0.343	0.343	0.142	0.142
95% CI lower bound	-0.066	-0.066	-0.027	-0.027
95% CI upper bound	0.189	0.189	0.185	0.185
N	415	415	407	407
Control Mean	-0.011	-0.011	-0.011	-0.011
Treatment Mean	-0.02	-0.02	-0.144	-0.144
R^2	0.754	0.754	0.787	0.787
Adj. R^2	0.741	0.741	0.776	0.776

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the student level. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. The following assessments are included in the index outcome for Rocketship: Dynamic Indicators of Basic Early Literacy Skills - DIBELS, and NWEA MAP test. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation.

WHAT'S NEXT

The research team is still analyzing the implementation and cost study data for 2023-24, as well as refining the analysis of impacts. Additionally, Rocketship's data will be pooled with other sites for the forthcoming personalized treatment effect analysis.

Rocketship Public Schools and BATA have continued partnering with PLI during school year 2024-25. The research team will report findings for the current school year once data collection and analysis are complete.

Preliminary findings from Winston-Salem/Forsyth County tutoring

THE PARTNERSHIP

In the 2023-24 school year, the PLI research team partnered with Winston-Salem/Forsyth County Schools in North Carolina, as well as a tutoring provider, University Instructors. This partnership was facilitated and supported by The Innovation Project, as with Guilford County Schools. We piloted two types of personalized learning interventions in three middle schools, and conducted an impact study. Tutoring began in February 2024 and lasted for 12 weeks.

THE INTERVENTIONS

In Winston-Salem, we piloted two personalized learning interventions—one more resource intensive, the other less so—both designed to provide mathematics support to middle school students. For both interventions, tutors interacted with students in person and provided instructional materials based on their students' unique needs. Students were grouped by tutoring coordinators. Tutoring took place in the classroom, with push-in tutors.

The key difference between the two interventions was:

- **HDT - 4:1 tutoring** - Tutors worked with four students at a time
- **SHDT - 8:1 tutoring with edtech** - Tutors worked with 8 students at a time, while leveraging the edtech platform iReady

Both interventions were intended to be delivered during students' class time for 30 minutes, three times a week, over the course of 10 weeks—an anticipated total of 750 minutes or 12.5 hours of math tutoring for the semester.

RESEARCH ACTIVITIES

The research team conducted three types of activities:

- **Randomization and impact analysis** - In both participating schools, most middle school students were eligible to receive either type of tutoring.[21] However, due to limited funding, not all students were able to be assigned to these interventions during the school day. We randomly assigned students to one of three conditions—4:1 tutoring, 8:1 tutoring with edtech, or a business as usual group—using classroom-level randomization. Because of this fair lottery, for a large enough sample, any differences in end-of-year test scores between the control and treatment groups were attributable directly to the interventions in expectation.
- **Surveys with tutors** - The PLI team obtained a roster of all tutors in PLI study schools in Winston-Salem/Forsyth County in the spring of 2024 (N = 16 tutors). Response rates were high with 93% of tutors responding.
- **Surveys and interview with coordinators** - A school staff member responsible for coordinating the tutoring program in each of the PLI study schools was surveyed and interviewed in the spring of 2024. The interviews touched on a variety of topics, with the questions designed to collect data about tutoring implementation and resource-use.

THE STUDENTS

As mentioned above, the majority of middle school students at the two schools were eligible for the two interventions. The randomized group included 1,569 students, with an even distribution of students across 6th, 7th, and 8th grade. Of the participants, approximately 91% were students of color, and 26% were English Language Learners. As shown in the below table, the randomized sample is well balanced on observable characteristics, with some slight imbalances for the SHDT sample.

[21] Typically, the few classes excluded from randomization were excluded due to being resource or other special needs classes where the intervention would be deemed disruptive.

Table 38: Baseline balance, Winston-Salem 2023-24, HDT

Covariate	Control Mean N = 581	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.27	-0.02	0.918	0.913	998
% Male	0.52	0	0.947	0.956	998
% Black	0.38	0.03	0.455	0.506	998
% Hispanic	0.48	-0.08	0.088*	0.14	998
% White	0.1	0.02	0.572	0.616	998
% Other	0.04	0.03	0.028**	0.05*	998
% Limited English Proficiency	0.27	-0.06	0.284	0.387	998
% Exceptional Children or 504 Plan	0.13	-0.02	0.753	0.781	998
Overall GPA	2.07	-0.02	0.911	0.929	976
Number of Attending Days	153.2	-0.37	0.85	0.881	979
Latest Available Math Score	0.01	0.09	0.579	0.622	1032
Second Latest Available Math Score	0.03	-0.05	0.782	0.604	966
Latest Available Reading Score	-0.03	0	0.997	0.997	1030
Second Latest Available Reading Score	0.04	-0.09	0.62	0.668	964
Grade 6	0.34	-0.07	0.531	0.596	1053
Grade 7	0.2	0.17	0.183	0.238	1053
Grade 8	0.45	-0.1	0.424	0.456	1053
***F*-Test - Baseline Cov.**			0***	0.415	1053

Note: The analysis sample is composed of HDT and BAU students randomized in a block in 2023-24 in Winston-Salem who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with HDT students are included. Reported p*-values test the difference in means for the HDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and only within the two relevant arms (e.g., HDT and BAU to test a hypothesis for HDT vs. BAU) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for the joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: **** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 39: Baseline balance, Winston-Salem 2023-24, SHDT

Covariate	Control Mean N = 581	Treatment Coefficient	*p*-value	RI *p*-value	N
Age	12.27	-0.34	0.127	0.212	953
% Male	0.52	-0.07	0.039**	0.086*	953
% Black	0.38	0.03	0.462	0.528	953
% Hispanic	0.48	0	0.98	0.983	953
% White	0.1	-0.05	0.07*	0.129	953
% Other	0.04	0.01	0.419	0.487	953
% Limited English Proficiency	0.27	0.03	0.657	0.686	953
% Exceptional Children or 504 Plan	0.13	-0.05	0.268	0.373	953
Overall GPA	2.07	-0.19	0.325	0.44	928
Number of Attending Days	153.2	-2.88	0.189	0.271	937
Latest Available Math Score	0.01	-0.15	0.276	0.361	977
Second Latest Available Math Score	0.03	-0.31	0.045**	0.091*	920
Latest Available Reading Score	-0.03	-0.17	0.128	0.196	972
Second Latest Available Reading Score	0.04	-0.35	0.007***	0.029**	921
Grade 6	0.34	0.02	0.895	0.911	999
Grade 7	0.2	0.3	0.02**	0.06*	999
Grade 8	0.45	-0.32	0.007***	0.037**	999
***F*-Test - Baseline Cov.**			0***	0.076*	999

Note: The analysis sample is composed of SHDT and BAU students randomized in a block in 2023-24 in Winston-Salem who have at least one end-of-year test score in their tutored subject. Only BAU students who were randomized in a block with SHDT students are included. Reported p*-values test the difference in means for the SHDT and BAU students in this sample. To conduct the pairwise tests, we regress the baseline covariate on a treatment indicator and randomization block fixed effects. No imputation was carried out and the number of observations vary reflecting availability of the variable, as shown in column "N". The latest and second latest available scores are defined as the most recent assessments available before randomization in each subject, standardized within grade, subject, school-year, and assessment.*

In the final row, we test the joint hypothesis of overall differences in baseline characteristics between the treatment and the control group. F*-tests are run using imputed values to account for missing data using a mean method within site, year, school, and grade. For any covariates that remain missing after the imputation procedure, cells are assigned a value of 0. To test the joint hypothesis, we regress a treatment indicator on baseline covariates, corresponding missingness indicators, and grade and randomization block fixed effects and calculate the resulting* F*-statistic from this regression. Missingness indicators are included in the regression model but not in the* F*-test.*

We report both the p*-value and randomization inference* p*-values, to avoid distributional assumptions. To calculate the randomization inference* p*-values, we randomly re-assign the treatment indicator within randomization blocks (fixing the randomization rate of each arm within the block) and estimate the corresponding test-statistic (*p*-value for pairwise tests,* p*-value of the* F*-statistic for joint test) from each placebo draw. We repeat this process 1,000 times. In the distribution of 1,000 placebo treatments, we see where the originally calculated test statistic lies, and report the percentile rank, which determines the RI* p*-value.* p*-values clustered at the randomization unit level are also reported and statistical significance is denoted as follows: *** p < 0.01, ** p < 0.05, * p < 0.10.*

TAKE-UP AND DOSAGE

In both the HDT and SHDT analytic samples, approximately 75% of students assigned to the treatment group received tutoring, defined as participating in at least one tutoring session. About 2% of control students for HDT and SHDT students also received tutoring. Students in both the HDT and SHDT groups each received, on average, about 13 sessions of the intervention, over the course of the study, conditional on attending at least one session, or about 400 minutes of tutoring (almost seven hours) over the course of the semester.

PRELIMINARY IMPACTS AND POTENTIAL INTERPRETATIONS

Our measure of student learning in this analysis is an index constructed from student scores on two end-of-year math assessments: a standardized state assessment, either the North Carolina End of Grade (EOG) or End of Course (EOC) assessment, depending on the student's mathematics grade level, and their iReady math score. We created this index by calculating the simple average of the two test scores, each standardized within grade, subject, and school year. As shown below, this primary outcome index is available for 94% of students.

Table 40: Take-up and dosage, Winston-Salem 2023-24

HDT Analysis Sample		
	BAU	HDT
N	581	472
% Received Treatment	1.72	76.48
Average Attended Sessions (Conditional)	7.5	13.45
Average Attended Sessions (Unconditional)	0.13	10.28
Average Scheduled Minutes (Conditional)	225	403.38
Average Scheduled Minutes (Unconditional)	3.87	308.52
% Students Missing Dosage Data	0	0

SHDT Analysis Sample		
	BAU	SHDT
N	581	418
% Received Treatment	1.72	74.88
Average Attended Sessions (Conditional)	5.4	13.14
Average Attended Sessions (Unconditional)	0.09	9.84
Average Scheduled Minutes (Conditional)	162	394.12
Average Scheduled Minutes (Unconditional)	2.79	295.12
% Students Missing Dosage Data	0	0

Note: The analysis sample is the sample of randomized students in Winston-Salem (see Table 1) that have a non-missing primary outcome measure. The primary outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject.

Dosage is calculated as of the latest assessment in the primary index outcome. For HDT students and their BAU counterparts, only HDT sessions are counted for take-up and dosage. For SHDT students and their BAU counterparts, only SHDT sessions are counted for take-up and dosage. Conditional dosage is the average dosage for students who received at least one tutoring session. Unconditional dosage is average dosage for all students in the analysis sample. If a control student was recorded as taking up a tutoring session, but the data does not allow us to discern if they attended an HDT or SHDT session, or they appear to have taken both HDT and SHDT, we leave their treatment indicators for both as zero. We do not directly observe minutes of dosage in our data. Instead, we approximate the number of minutes attended by multiplying the number of sessions each student attended—which we observe at the student level—by the scheduled length of those sessions, which varies by site and randomization block.

To estimate program impacts on our primary outcome, we compute both the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects, following the analysis decisions outlined in our pre-registered analysis plan (see Appendix 1). The ITT estimate captures the effect of being offered the opportunity to participate in tutoring, while the TOT estimate captures the effect of actual participation, defined as attending at least one tutoring session.

We did not find statistically significant impacts of HDT or SHDT on student learning as measured by an index of end-of-year math scores at Winston-Salem.

On average, students who received at least one HDT or SHDT session gained an additional 0.028 and 0.035 standard deviations, respectively, in end-of-year math scores compared to their peers who did not receive either type of tutoring. However, we cannot reject the null hypothesis of no effect for these treatments. We hypothesize that this null result is due to low dosage—about seven hours of math tutoring, about a third of the designed dosage—over a short 12 week period. Additionally, while the tutoring model was designed as a push-in model, most tutors (64%) reported that sessions were most commonly hosted outside of the classroom ("pull-out instruction"). Finally, SHDT tutors reported issues with accessing iReady during their sessions.

Table 41: Differential attrition, Winston-Salem 2023-24

Sample	% Students with Outcome Missing	% Treatment Students with Outcome Missing	% Control Students with Outcome Missing	*p*-value
HDT	6.07%	5.03%	6.89%	0.412
SHDT	6.81%	6.7%	6.89%	0.678

Note: The outcome is a simple average of all available end-of-year, standardized tests (relative to the control group score distribution within grade) a given student takes in the tutored subject. The following assessments are included in the index outcome for Winston-Salem: North Carolina End-of-Grade (EOG) and End-of-Course (EOC) state assessments and iReady. Only students who are missing all assessments are missing an outcome; if a student takes any end-of-year assessment in the tutored subject, they have an outcome. We generate p*-values by regressing outcome missingness on treatment status and all covariates shown in the balance tables along with grade and randomization block fixed effects. We impute missing values for baseline variables at the level of year of randomization, school, and grade. Missingness by treatment status and assessment is available upon request.*

Table 42: Impact estimates, Winston-Salem 2023-24

	HDT		SHDT	
	ITT	TOT	ITT	TOT
Estimate	0.021	0.028	0.027	0.035
Std. Error	0.068	0.091	0.048	0.062
***p*-value**	0.761	0.764	0.576	0.575
95% CI lower bound	-0.112	-0.156	-0.067	-0.090
95% CI upper bound	0.153	0.211	0.120	0.160
N	1053	1053	999	999
Control Mean	-0.007	-.007	-0.007	-0.007
Treatment Mean	0.046	0.046	-0.232	-0.232
R^2	0.633	0.632	0.640	0.640
Adj. R^2	0.621	0.620	0.627	0.627

Note: This model includes all covariates shown in the balance tables along with missingness indicators, grade indicators, and randomization block fixed effects. Standard errors are clustered at the level of randomization. We impute missing values for control variables at the level of year of randomization, school, and grade. The outcome of interest is an index of all available relevant EOY standardized test scores in the tutored subject. We standardized test scores at the level of year of randomization, site, assessment, and grade using the control mean and standard deviation. The following assessments are included in the index outcome for Winston-Salem: North Carolina End-of-Grade (EOG) and End-of-Course (EOC) state assessments and iReady.

WHAT'S NEXT

2023-24 was the only year of PLI participation for Winston-Salem/Forsyth County Schools, so no new data is forthcoming for this site. However, the research team is still analyzing the implementation and cost study data for 2023-24, as well as conducting additional impact analysis on secondary outcomes. In addition, this data will be pooled with other sites for the forthcoming personalized treatment effect analysis.

Appendix I: Pre-Analysis Plans

The pre-analysis plan for this study is posted on Open Science Framework at the link below.

https://osf.io/fkjmn/

Appendix II: Implementation and Cost Data Collection Methodology

BACKGROUND ON THE IMPLEMENTATION AND COST STUDY

The priorities of the Personalized Learning Initiative's (PLI) implementation research study have iterated year over year in response to the developments of the broader PLI study and in response to site needs. Beginning in 2021-22, the PLI implementation research study piloted qualitative data collection in Chicago Public Schools with tutors and school leaders to identify facilitators and barriers to successful launch and implementation of the Tutor Corps program. The insights of the pilot year informed development of data collection plans for 2022-23 in Chicago Public Schools, Fulton County Schools and New Mexico Public Education Department. This second year of the study focused on understanding facilitators and barriers to student receipt of intended dosage of tutoring, identifying key resources needed to deliver tutoring as intended, and learning about key school stakeholders' perception of the tutor program quality. Data collection in 2022-23 included surveys of tutors and tutor coordinators as well as tutoring observation and interviews with school tutor coordinators, teachers and tutors.

Data collection and analysis from the first two years of the PLI study highlighted the importance of building strong support systems at local education agencies and within schools to monitor and manage provision of tutoring as intended. These insights have been summarized in a checklist for education leaders starting tutoring programs, a research brief, a commentary for district administrators and a podcast. These insights also informed the 2023-24 PLI study's development of a more robust technical assistance offering to study sites to support implementation as intended, the design of the 2023-24 cost study, and efficient data collection strategies to systematically document tutor program implementation at scale.

AIMS FOR IMPLEMENTATION AND COST RESEARCH IN 2023-24

In 2023-24, the PLI study scaled to sites affiliated with eight education agencies and 84 schools with considerable variation in the types of tutoring programs being offered and evaluated.[22] Additionally, the PLI research team expanded to include a dedicated cost study team with distinct but complementary research aims to the implementation research study. As a result, the cost and implementation teams collaborated on data collection to explore program implementation and resource use. Aims and data collection strategies of both teams are described below:

Implementation. Given the diversity of tutoring program models in 2023-24, the implementation study prioritized systematic documentation of how tutoring programs were designed and implemented in 2023-24. Additionally, the team leveraged insights from prior years to design more targeted and efficient means of collecting information on facilitators and barriers to implementation as designed and initiated a line of inquiry into tutors' career aspirations. Finally, the team bolstered efforts to document the extent to which tutoring was being implemented alongside other supplemental personalized learning interventions for students. Implementation research team is investigating the following questions for 2023-24:

- **Content:** What are the characteristics of tutoring programs designed and implemented in the Personalized Learning Initiative?
    - To what extent were tutoring programs implemented as intended?
    - To what extent did PLI students receive tutoring programs at the intended dosage?
    - Who are the tutors working in PLI study schools?

[22] There were 84 school sites participating in data collection for 2023-24. Of these 84 schools, 83 are included in the analytic sample for the impact analysis.

- **Contexts:** What challenges did schools face in implementing tutoring programs as designed? What factors facilitated implementation as intended?
    - How do tutors and coordinators perceive the quality of the school's tutoring program?
    - To what extent do tutors plan to remain in the K–12 education sector?
- **Contrast:** What other types of personalized learning experiences were provided by study schools alongside tutoring?

Cost. 2023-24 represented the first concerted data collection on cost and resource use. The cost team explored the costs of the 2023-24 tutoring programs both as designed and as they were eventually implemented. In assessing costs as designed, specified program parameters were inventoried and assigned costs to ensure that the programs intended to be less expensive were genuinely less expensive and to assess the magnitude by which they differed. This analysis informed the design of the thorough analysis of the tutoring programs as implemented. The cost team analysis covers the following research questions, with analysis ongoing:

- **Design:** What are the costs of PLI tutoring programs as *designed?* How do intended HDT and SHDT program costs differ?
- **Implementation:** What are the costs of PLI programs as *implemented?* How do actual HDT and SHDT program costs differ?
- **Cost-Effectiveness:** What is the cost-effectiveness of PLI programs as implemented? How does the cost-effectiveness ratio of HDT compare with SHDT? How do these compare with previously studied models of high dosage tutoring?

Further inquiries on cost differentials across different PLI domains, such as differences in costs between the following dichotomies are of interest and will be investigated if warranted.

- Virtual tutors vs. in-person tutors
- Intervention run by school system vs. vendor
- Whether edtech was integrated
- By tutor qualifications

IMPLEMENTATION AND COST DATA COLLECTION

Table II.1 summarizes key constructs measured by the data collection activities used to address the implementation and cost research questions. Table II.2 provides information on the sampling strategy and response rate for the primary data collection activities.

Table II.1: Data collection activities

Data collection activity	Key constructs measured
Fall 2023 tutor survey	**Content** • Tutors report of student attendance on a typical day (for the purposes of triangulating information from administrative records on attendance) • Tutor background characteristics and prior work experience
Spring 2024 tutor survey	**Content** • Tutor report on typical program structure (as implemented) • Tutor report on student attendance on a typical day • Tutor background characteristics and prior experience (new respondents only) **Context** • Tutor future career aspirations • Tutor perception of student engagement and behavior • Tutor report on quality of their experience and confidence in their abilities • Tutor report on factors contributing to student attendance and session cancellation **Cost** • Tutor time use • Tutor training • Tutor's use of their own personal resources for tutoring • Tutor qualifications
Spring 2024 school coordinator survey	**Content** • School coordinator role • School coordinator time-use • School coordinator report on tutoring program set-up including resource use **Context** • Coordinator perceptions of tutor quality • Coordinator report on strengths and challenges of tutor program implementation • Coordinators perception of key elements of school context **Contrast** • Reach of other supplemental personalized learning supports at the school **Cost** • Coordinator and tutor time use • Coordinator and tutor training • Other involved personnel • Facility, supply, and material use

Table II.1: Data collection activities, continued

Data collection activity	Key constructs measured
Spring school coordinator interview	**Content, Contrast, Context** • More details on survey topics including how coordinators made decisions about resource allocation & perception of facilitators/barriers to implementation **Cost** • Coordinator and tutor time use • Coordinator and tutor training • Other involved personnel • Facility, supply, and material use • Unexpected resource use for day to day program operation, to respond to issues, or both • Distribution of resources between SHDT and HDT programs, when relevant
Student attendance data	• Varies by site but all sites permit a count of the number of sessions attended, by student
PLI technical assistance team notes	• Systematic observations of tutoring sessions • Notes from structured conversations with site coordinators about both programs as designed and as intended, including costs and resource use

Table II.2: Sampling strategy and response rates

Primary data collection activity	Sampling strategy	Response rate
Fall 2023 tutor survey	All tutors on roster in early Fall 2023 at the four sites implementing tutoring for full Fall 2024 semester (N = 396)	87.9%
Spring tutoring survey	All tutors on roster in early Spring 2024 in the eight sites implementing tutoring for full Fall 2024 semester (N = 467)	89.7%
Spring 2024 school coordinator survey	All school staff identified as coordinator of the tutoring program(s) at their school in Spring 2024 (N = 84 school staff)	90.5%
Spring school coordinator interview	Two or more schools per program type were purposively sampled with the intention to capture a range of program implementation styles and qualities with a range of school characteristics (N = 94 coordinators invited to an interview)	75.53%

Addressing variation. Given the number and diversity of sites and the preponderance of tutoring programs covered in the 2023-24 PLI study, it was important to implementation and cost study research aims to be able to define the main domains on which program design and implementation varied. The impact study used hundreds of study random assignment blocks, which was far too granular a unit of analysis for an implementation or cost study to document implementation. On the other side of the coin, site level analysis across the eight sites would mask within site variation, which was substantial.

To bridge the gap between randomization blocks and sites, the cost and implementation research teams specified a set of 25 program types along the following domains:

- Site
- Grade level
- Subject of tutoring
- Study's HDT/SHDT designation (e.g., typically defined by tutor-to-student ratio and integration of edtech)
- When a meaningful parameter of variation within site: curricular materials and/or classroom setting (e.g., type of math class that tutoring was delivered in)

As shown in the figure below, the program types were designed to map upwards to permit aggregation for site level summaries and downwards to support generalization to RA blocks. Some schools in the study included multiple program types within them. Thus, school coordinator interviews and surveys asked coordinators to report separately about each program type.[23] Additionally, technical assistance teams were asked to indicate program type when documenting observations and issues encountered with each school's program.

Tutor program types, shown here in blue, provided a bridge between random assignment blocks (often at the class or period level) and the site level.

Figure II.A: Example of program mapping

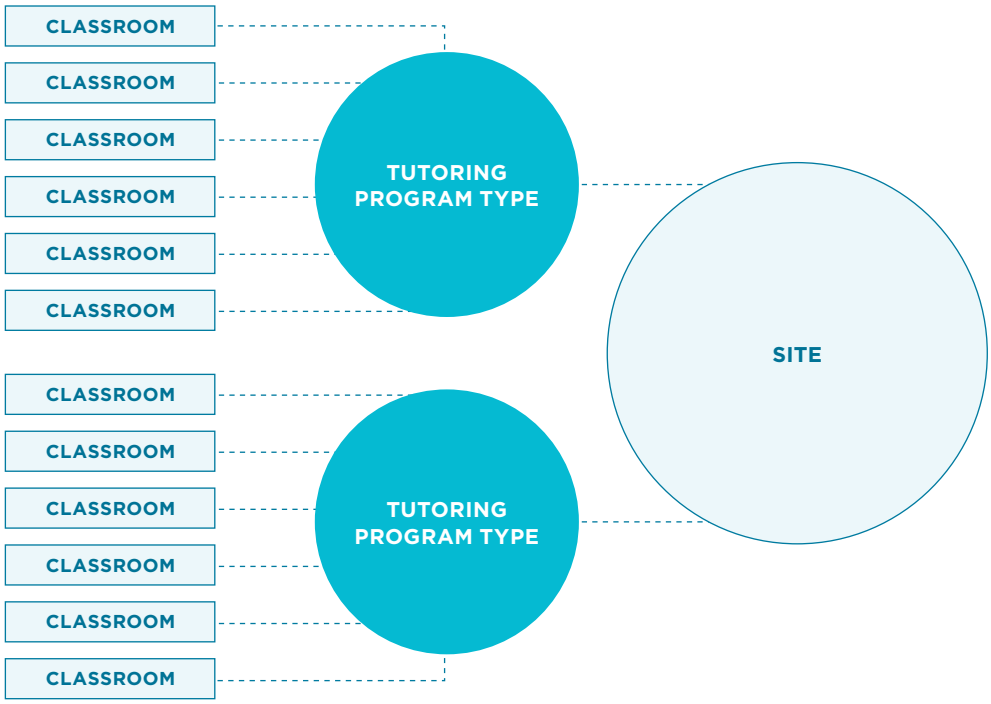


Table II.3: Program types in 2023-24 PLI sites

Program types in 2023-24 PLI sites					
Site	**School type**	**Subject**	**Study assignment**	**Program type**	**Notes on format differentia**
Chicago	K-8	ELA	HDT	C1. Elem Lit HDT	
		Math	HDT	C2. MS Math HDT	
			SHDT	C3. MS Math SHDT	
	HS	Math	HDT	C4. HS Math HDT	
			SHDT	C5. HS Math SHDT	
Fulton	Elem	ELA	HDT	~~F1. Elem ELA HDT~~	~~Tutoring without Lightning Squad TA support~~
				F2. Elem ELA HDT w LS	Tutoring with Lightning Squad TA support
			SHDT	~~F3. Elem ELA SHDT~~	~~Tutoring without Lightning Squad TA support~~
				F4. Elem ELA SHDT w LS	Tutoring with Lightning Squad TA support
		Math	HDT	F5. Elem Math HDT	
			SHDT	F6. Elem Math SHDT	
	MS	Math	HDT	F7. MS Math, HDT	
			SHDT	F8. MS Math, SHDT	
	HS	Math	HDT	F9. HS Math, HDT	
			SHDT	F10. HS Math SHDT	
Greenville	GS	Math	HDT	GV1. MS Math HDT	
			SHDT	GV2. MS Math SHDT	
Miami	MS	Math	HDT	M1. MS Math Foundational HDT	Tutoring in foundational math class
				M2. MS Math Core HDT	Tutoring in core math class
			SHDT	M3. MS Math SHDT	
North Carolina	MS	Math	HDT	NC1. Guilford HDT	Guilford
			SHDT	NC2. Guilford SHDT	
			HDT	NC3. Winston Salem HDT	Winston Salem
			SHDT	NC4. Winston Salem SHDT	
New Mexico	MS	Math	HDT	NM. MS Math HDT	
Rocketship	Elem	ELA	HDT	R1. ELA HDT	
			SHDT	R2. ELA SHDT	

Note: All schools in program types 1 and 3 in Fulton eventually took up Lightning Squad, warranting the collapsing of the non-Lightning Squad versions into the Lightning Squad version.

ANALYTIC APPROACH: IMPLEMENTATION RESEARCH

To support sharing early insights with sites, spring 2024 survey data from tutors and coordinators was summarized by site to document how programs were implemented (content), tutors' and coordinators' perceptions of program quality and the quality of their experience (context), the reach of other personalized learning supports at the school alongside tutoring (contrast) and to describe tutors' background characteristics, credentials and career aspirations. Additionally, administrative data on the number of sessions attended by each student was summarized to describe dosage and assess the extent to which students received the intended dosage intended.

To achieve the broader aims of the implementation research team, the team is also summarizing information by tutoring program type about the design and implementation of key program features. Due to the diverse designs of tutoring program models across and within sites during 2023-24, traditional methods of developing indices to summarize the implementation fidelity of all intervention core components for all programs in the study were not very useful. Instead, the implementation research team and the cost study team have worked together to identify a comprehensive list of the program features relevant to all tutor program types including student dosage, tutor-to-student ratio, tutoring location, when in the school day students receive tutoring, use of edtech, and instructional materials. For each program feature, the implementation and cost teams identified the most reliable data source (e.g., survey data, interview data with coordinator or technical assistance team notes) for documenting the design and implementation of the feature. Information on design and implementation is not available for all schools for each program feature. As a result the implementation research team is relying on the cost study team's compilation of this information for a subset of case study schools within each program type (see description below on the cost study team's sampling methods). Thus, each of the 25 program profiles will provide information on the intended program type design and implementation based on information collected from each program type's case study schools.

Analysis of the individuals profiles will permit rich description for policymakers and practitioners about what tutoring programs look like in practice and will support contextualization of cost findings by program type. Summaries of program profiles at the site level and overall will provide context for impact findings at those levels.

ANALYTIC APPROACH: COSTS

Costs *as designed methodological approach.* The research team generally followed the principles of the ingredients method (Levin et al. 2017) to assess the first cost research question: "What are the costs of PLI tutoring programs as designed? How do intended HDT and SHDT program costs differ?" Following the ingredients method, the resources that comprise a given tutoring program were listed along with the quantities with which they were used and any notes on specific resource's qualities or defining characteristics.

The ingredients method typically involves a thorough data collection effort to account for all of the resources needed to run a program. However, given that the as-designed cost estimates were meant to reflect the program costs as was originally planned, not how it was actually implemented in schools, such data collection was not possible. Instead, we needed to gather data about resources and their quantities that documented the programs' intended design. This analysis involved reviewing program design documents. The team also conducted interviews with site teams to gain additional details about the planned programs. It is important to note that this process excludes school-level nuance or any variation or adjustment that might have occurred after the implementation began.

Costs as implemented methodological approach. While costs as designed inherently deviated from the ingredients method due to the nature of the inquiry, the costs as implemented inquiry summarized in the second cost research question more closely adhered to the ingredients method to investigate "what are the costs of PLI programs as implemented? How do actual HDT and SHDT program costs differ?"

The costs as implemented analysis proceeded in the following steps:

1. **Sampling.** The cost team worked with site teams to identify schools for the cost analysis. These selections represented a purposive sample and sought to represent a range of characteristics within program type. These characteristics were as perceived by the site teams and included:

   a. School characteristics (size, urbanicity, format, leadership style)

   b. Implementation enthusiasm and perceived implementation quality

Additionally, schools with insufficient data with which to estimate costs were avoided. The initial cost sample goal was three schools per program type, however, in a number of program types, there were one or two schools total, inherently limiting the sample to the total number of schools. In program types where three schools were available, one school was selected as an alternate and the other two were used for primary analysis.

The cost analysis sample appears in Table II.4 below.

The cost sample included 46 school by program type combinations, with 38% of the school by program type combinations represented in the cost sample and an average of 1.84 schools per program type.

Table II.4: Cost sample across program types

Site	Program type	Schools in cost sample	Schools in study	Sample proportion
Chicago	C1. Elem Lit HDT	2	21	10%
Chicago	C2. MS Math HDT	2	14	14%
Chicago	C3. MS Math SHDT	1	3	33%
Chicago	C4. HS Math HDT	1	1	100%
Chicago	C5. HS Math SHDT	2	7	29%
Fulton	F2. Elem ELA HDT w LS	2	3	67%
Fulton	F4. Elem ELA SHDT w LS	2	3	67%
Fulton	F5. Elem Math HDT	2	4	50%
Fulton	F6. Elem Math SHDT	2	4	50%
Fulton	F7. MS Math, HDT	2	3	67%
Fulton	F8. MS Math, SHDT	2	3	67%
Fulton	F9. HS Math, HDT	1	1	100%
Fulton	F10. HS Math SHDT	1	1	100%
Greenville	G1. MS Math HDT	2	3	67%
Greenville	G2. MS Math SHDT	2	3	67%
Miami	M1. MS Math Foundational HDT	2	7	29%
Miami	M2. MS Math Core HDT	1	5	20%
Miami	M3. MS Math SHDT	1	1	100%
North Carolina	N1. Guilford HDT	2	2	100%
North Carolina	N2. Guilford SHDT	2	2	100%
North Carolina	N3. Winston Salem HDT	2	3	67%
North Carolina	N4. Winston Salem SHDT	3	3	100%
New Mexico	NM. MS Math HDT	3	19	16%
Rocketship	R1. ELA HDT	2	2	100%
Rocketship	R2. ELA SHDT	2	2	100%
	Total	**46**	**120**	**38%**

2 **Data compilation.** The cost study team compiled data on the selected schools from the data sources listed earlier in this appendix as well as from program documentation and study records. This phase focused on recording information on the resources used, regardless of who paid for them, a description of the resources, and the quantity in which it was used.

3 **Pricing.** Each resource was then matched with a standardized national price that most closely reflected the anticipated market value and description of the resource and adjusted to 2023 US dollars. Standardized prices were drawn from national sources, such as the Bureau of Labor Statistics 2023 wage data and national retailers. There are three advantages to using standardized prices rather than the actual prices paid by sites.

a. *Reduced regional bias.* Cost differences due to regional economic conditions are substantially reduced so that the resulting estimates are comparable. Otherwise, the analysis risks identifying a difference between two tutoring programs that is due to regional price differences rather than actual program differences.

b. *Mitigating site-specific noise.* By using site level accounting data, the analysis would risk picking up nuance from the sites such as a discount offered through a local business partnership, items donated from a nonprofit, or reallocated resources. This sort of site level noise is not generalizable to another site.

c. *Enhanced generalizability for replication*. By addressing regional bias and limiting site-specific noise, the resulting estimates are most likely to be representative of the cost one might expect should the program be replicated at an additional site in any new location. Such replicability and external generalizability are critical to a study done at a national scale.

Tutor pricing. In *most* PLI sites, the intended tutor qualifications were minimal. Most sites had a vague goal of recruiting college students, community members, paraprofessionals, and other individuals. To apply a standard wage that was not subject to local market variation, we used the BLS median tutor wage for tutors working in school settings (Levin et al. 2017). After applying a methodologically standard adjustment (Shand and Bowden 2022) to reflect minimal HR costs related to paying staff, this wage equals $24 in 2023 dollars. Based on our available site level wage data, this wage is well situated in the range of tutor compensation, as seen below.

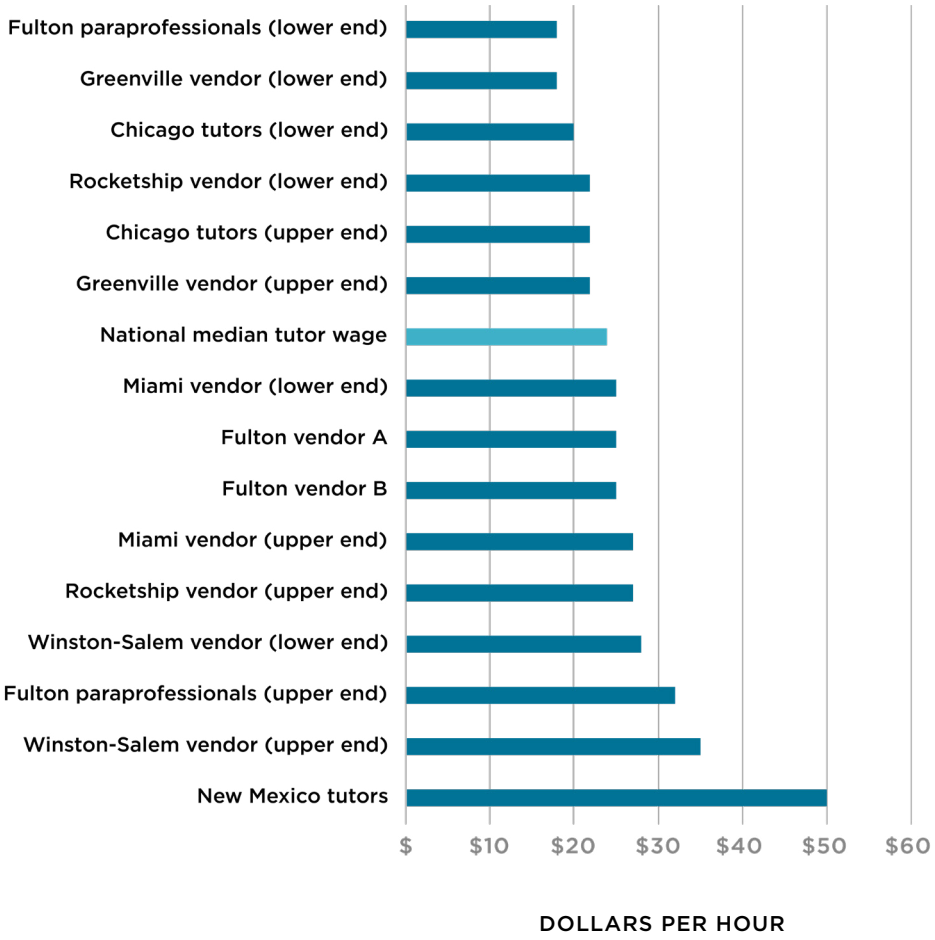
New Mexico was an obvious outlier in the wage distribution. In New Mexico, the high wage was an intentional part of the program design, intended to attract and retain a more qualified tutoring workforce. Here, the standard wage did not accurately reflect the value of the tutoring services likely to be recruited for the substantially higher wage, so in New Mexico cost estimates, the $50 wage was assumed to be the best wage for that specific program design. Sensitivity testing on this decision is ongoing.

4 **Calculating total costs.** After pricing each resource and multiplying by the observed quantities, all resources were summed to estimate the per school total price. This total was then divided by the number of students in the program to estimate a per pupil cost. Consistent with the calculations of the treatment-on-the-treated impact estimate, the cost team used the number of students that ever received a single session or more of tutoring in 2023-24. Costs were then adjusted to a standard program period of 30 weeks so that further analysis can isolate the length of the program from other cost drivers.

5 **Sensitivity analysis.** The cost estimates will subsequently be probed to understand the sensitivity of final analyses to analytical decisions by varying inputs where uncertainty exists around the quantity of a resource, the appropriate price for the resource, and any assumptions made.

Following the completion of cost and impact estimates, a cost-effectiveness or other further economic analyses such as a benefit-cost analysis may be conducted.

Figure II.B: Tutor wages across PLI sites, relative to national median, 2023-24



Note: All wages in 2023 USD. Some tutor wages were available as ranges; in that case the high and low end of the range are both presented in the figure.

Source: National wage is the Bureau of Labor Statistics 2023 median tutor wage plus a small multiplier to reflect employment costs, other wages were sourced from PLI site teams, employment advertisements, district pay tables, and vendor websites.

Appendix III: Sample ‘Participation Report’

On the following pages, we present a sample ‘Participation Report’ that illustrates take-up and dosage. This type of report was regularly produced for our PLI partners throughout the 2023–24 school year (typically every 4–10 weeks), and was used to inform ongoing technical assistance throughout the year.

Please note that the report below contains dummy data and does not reflect real participant information. We hope it is helpful for illustrative purposes.

PERSONALIZED LEARNING INITIATIVE

Tutoring Monitoring Report

PLI District

Monitoring Period: July 31, 2023 - September 01, 2023

Date Generated: September 01, 2023

This report is designed to provide information on PLI District's implementation of high dosage tutoring as part of the Personalized Learning Initiative (PLI). We hope this information can be helpful to you as you consider how tutoring is serving students throughout the school year. This document is not meant to be evaluative. Instead, its goal is to help aid in reflection on program implementation.

Please be advised that this monitoring report may undergo modifications in future months as we work to make it as useful as possible for schools and districts.

Executive summary

[Site teams will write an executive summary for the district-view report. No executive summary will be provided for school-view reports due to volume.]

How many students/schools were expected to receive tutoring in a given week?

Fig. 1. Number of Students Expected to Receive Tutoring Each Week

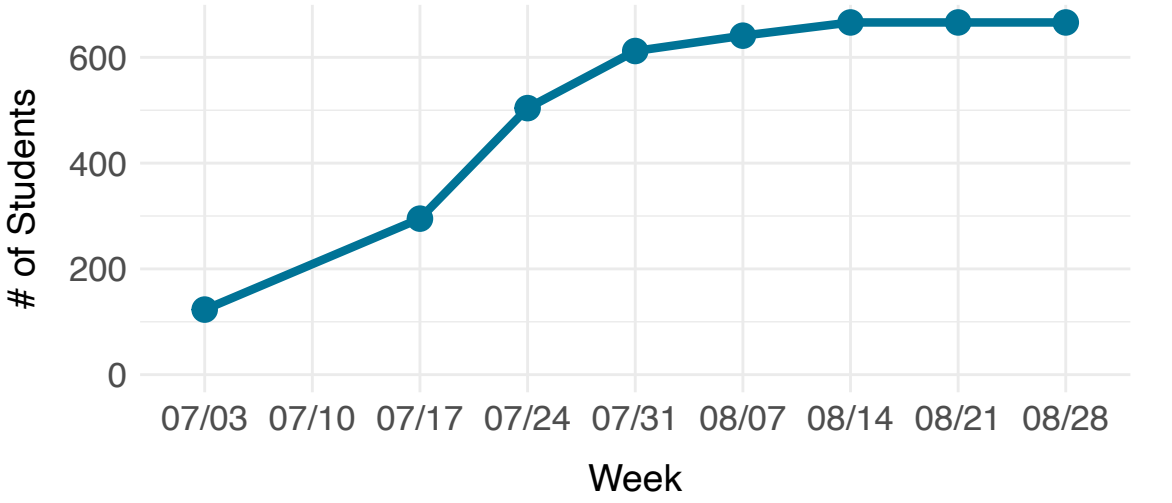
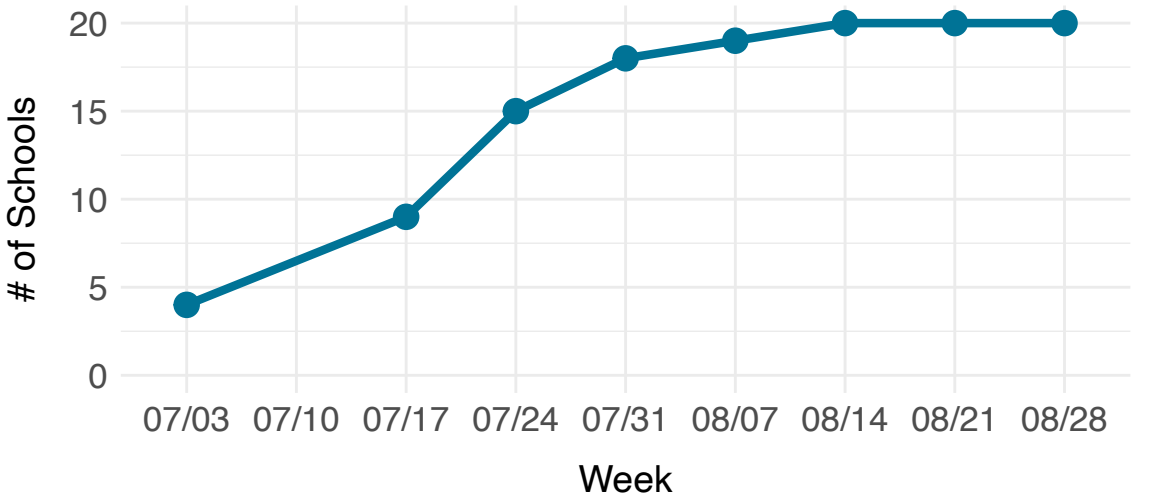


Fig. 2. Number of Schools Expected to Receive Tutoring Each Week



How many students scheduled for tutoring received tutoring between 07/31 and 09/01?

Tutoring Participation by Research Group

The table below displays the number of students receiving (or not receiving) tutoring in each assigned research group. Numbers highlighted in red indicate a difference between the intended and actual number of students receiving tutoring.

Table 1: Participation by Treatment Assignment (Number of Students)

	Received Tutoring This Period	
Treatment Assignment	No	Yes
BAU	334	0
HDT	25	206
SHDT	117	318

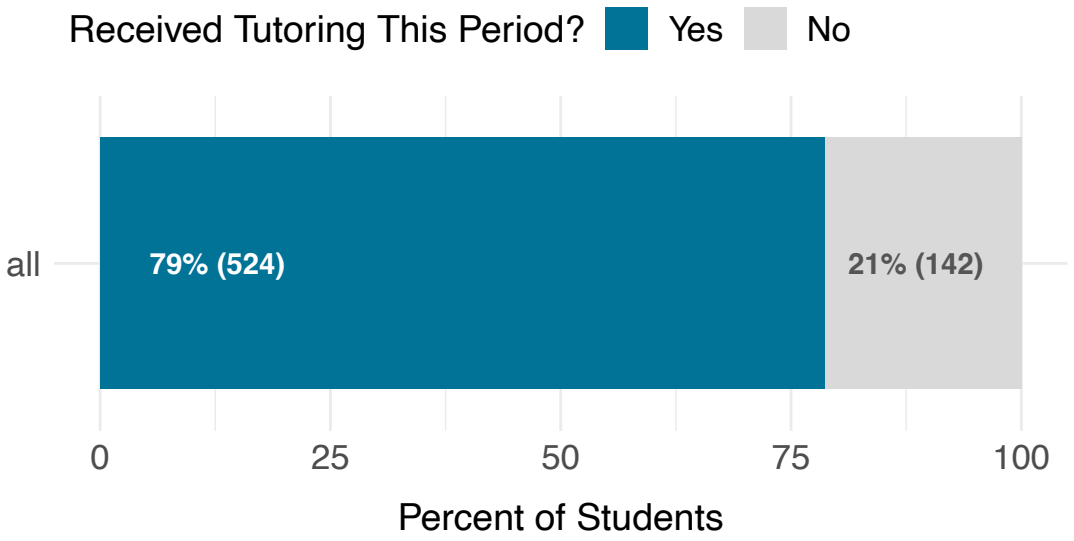
Note: Participation and dosage numbers on subsequent pages are limited to students who were assigned to tutoring in the PLI study. Students who have attended tutoring sessions but who were not assigned to tutoring will be excluded from the analysis.

Tutoring Participation and Data Availability

Table 2: Breakdown of tutoring participation during the monitoring period, among students scheduled for tutoring

Category	Number of Students	% of Scheduled Tutoring Students
Attended at least one session	524	78.7
No attendance record	142	21.3
Total	666	100.0

Fig. 3. Overall Participation Among Students Scheduled for Tutoring



524 students received at least one tutoring session during the monitoring period (07/31-09/01). This represents 79% of students scheduled for tutoring in the PLI study.

Participation is defined as whether a student had a present attendance record for at least one tutoring session during the monitoring period. Participation data is limited to the 666 students who were scheduled for tutoring.

Fig. 4. Participation by Subject

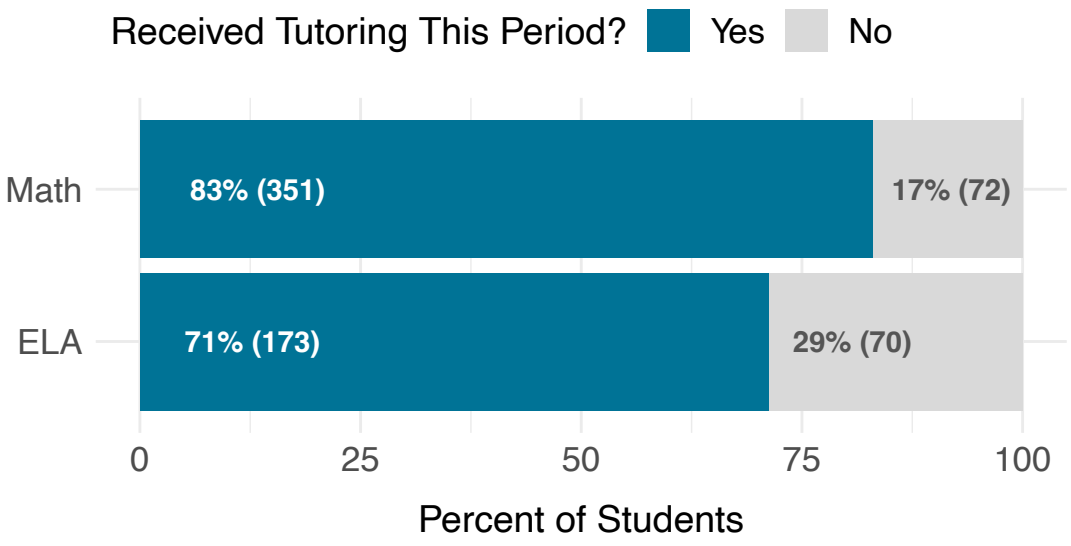
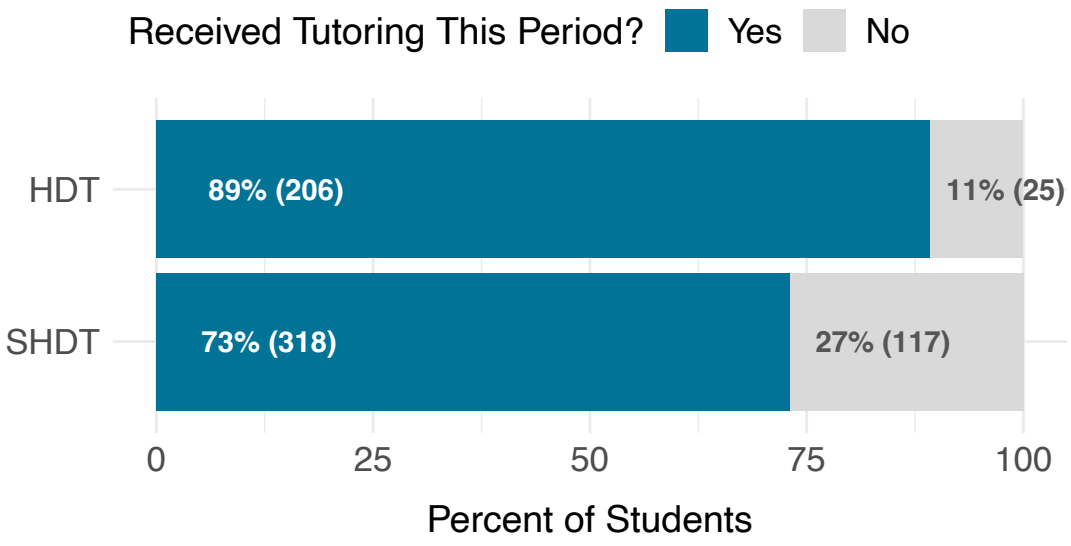


Fig. 5. Participation by Assigned Tutoring Type



Summary of Participation Rates, between 07/31 and 09/01

The tables below show the share of students in each grade, tutoring type, and subject who attended at least one tutoring session during the monitoring period

Table 3: Participation Rates for Math

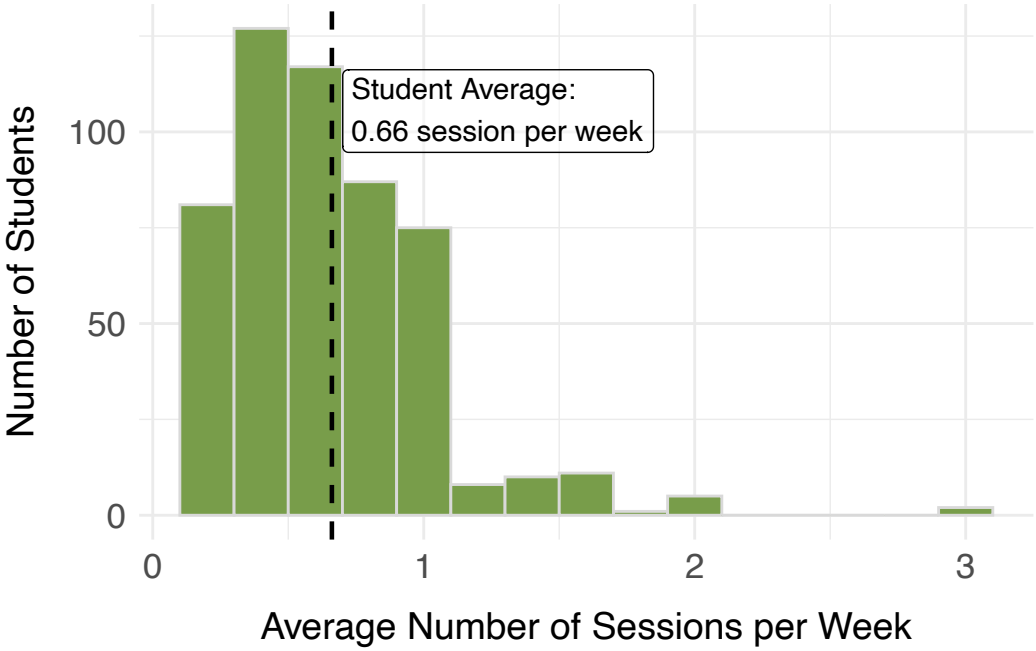
Grade	HDT	SHDT	All Tutoring Types
Grade 6	91%	80%	83%
Grade 7	95%	82%	85%
Grade 8	100%	62%	74%
All Grade Levels	94%	80%	83%

Table 4: Participation Rates for ELA

Grade	HDT	SHDT	All Tutoring Types
Grade 6	94%	41%	71%
Grade 7	79%	53%	69%
Grade 8	/	100%	100%
All Grade Levels	86%	51%	71%

How much tutoring did students receive between 07/31 and 09/01?

Fig. 6. Sessions per Week among Students Receiving Tutoring



Tutoring dosage is limited to the 524 students scheduled for tutoring who received at least one tutoring session during the monitoring period.

Within each school, dosage is measured between that school's tutoring start date and 09/01/2023 . For example, if the monitoring start date is 09/01 and a school began tutoring on 09/14, dosage will be measured in that school starting from 09/14.

Fig. 7. Number of Sessions per Week by Research Group & Subject

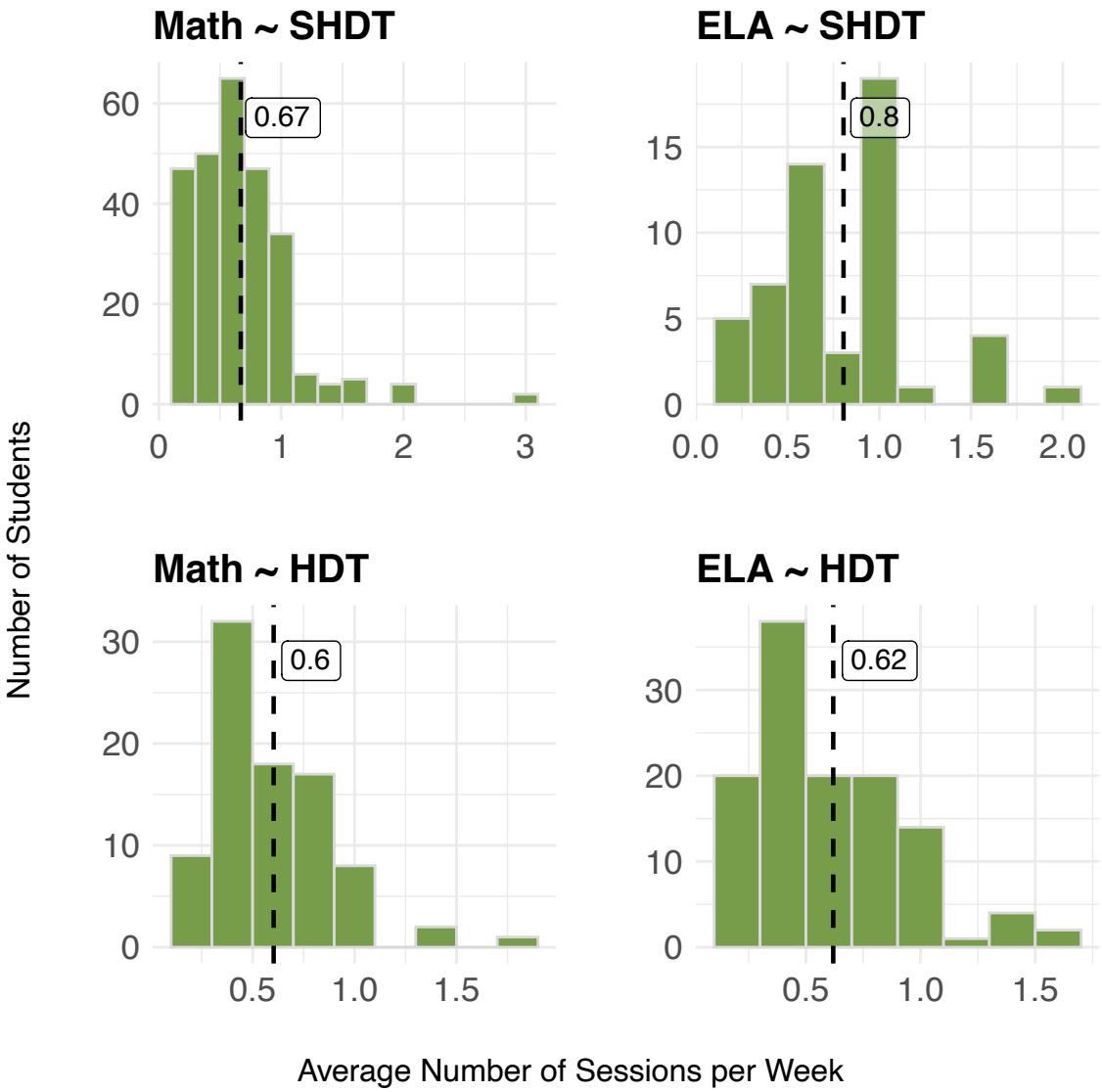
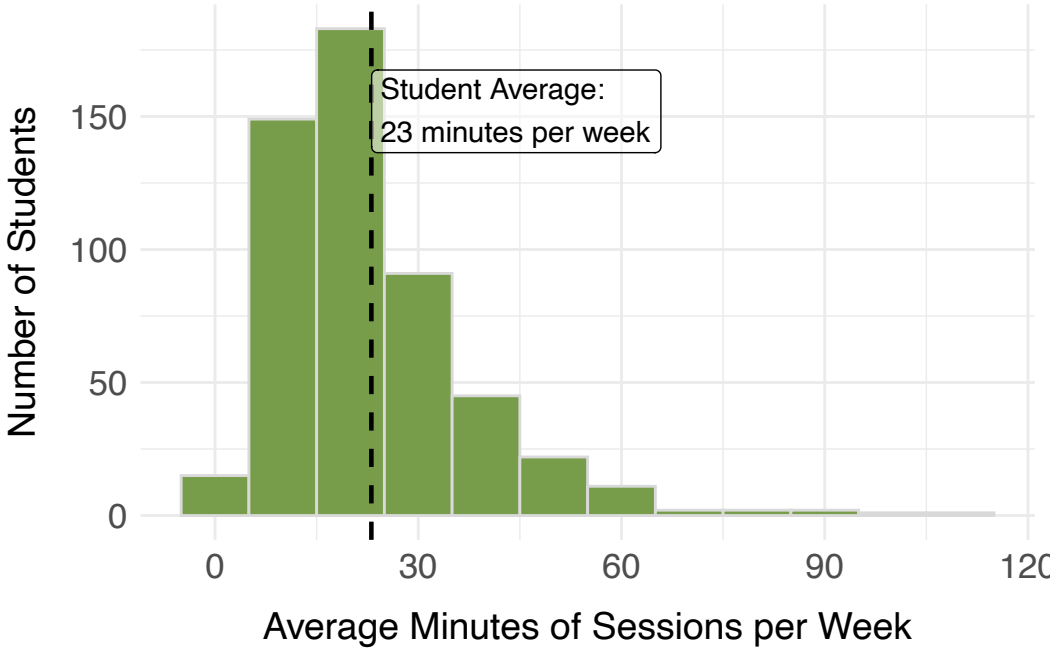


Fig. 8. Distribution of Average Minutes per Week



Tutoring dosage is limited to the 524 students scheduled for tutoring who received at least one tutoring session during the monitoring period. Plot reflects the average minutes of tutoring undergone by each student per week during the monitoring period.

Within each school, dosage is measured between that school's tutoring start date and 09/01/2023 . For example, if the monitoring start date is 09/01 and a school began tutoring on 09/14, dosage will be measured in that school starting from 09/14.

How have tutoring participation and dosage changed over time?

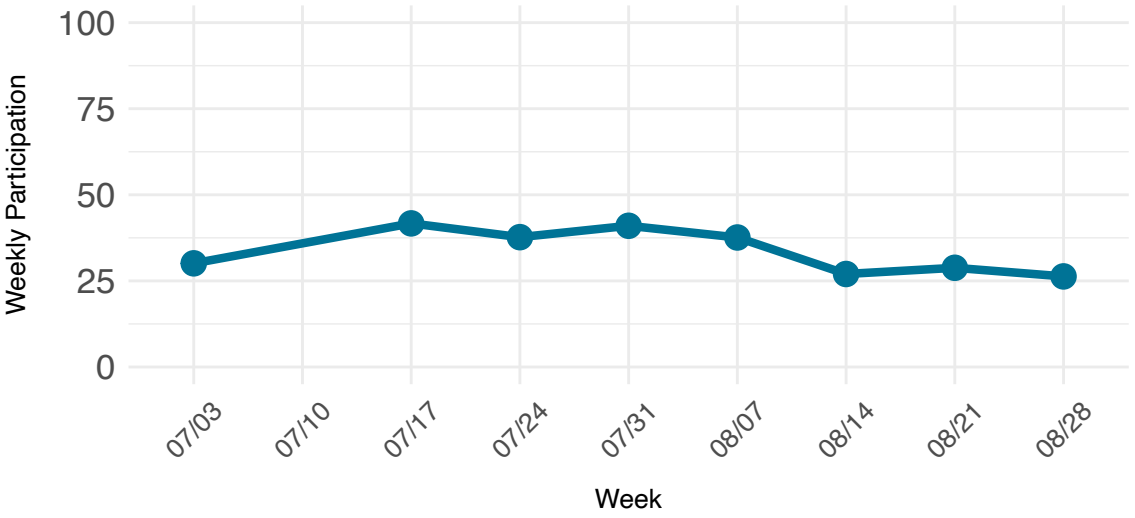
Participation this Monitoring Period vs Prior Period

Table 5: Change in Student Participation

Student Status	Number of Students
Never received tutoring in this period or prior period	82
Newly received tutoring this period	306
Received tutoring in the prior period, not this period	60
Received tutoring in this period and prior period	218

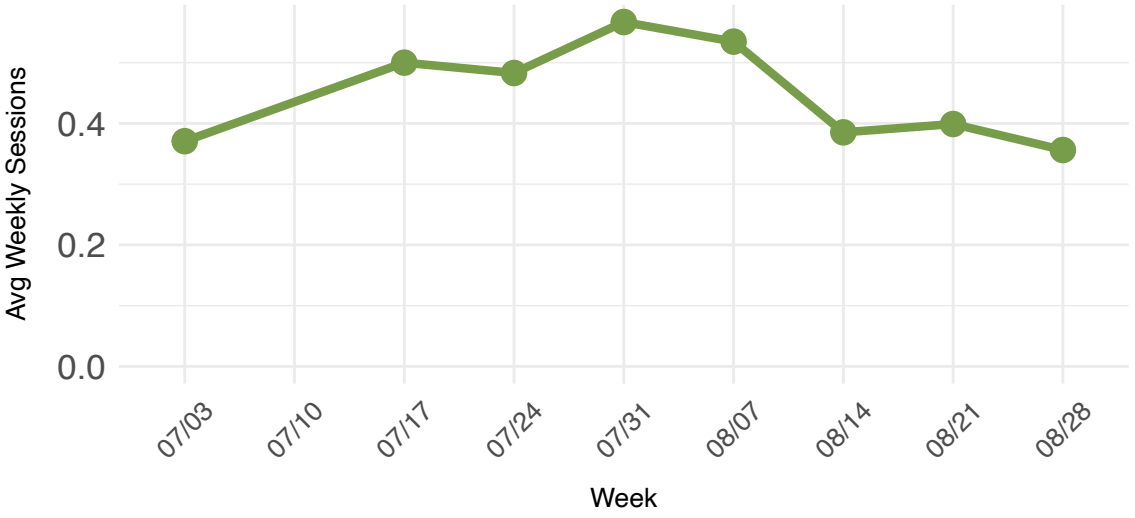
Note: Participation is defined as students who attended at least one tutoring session during the monitoring period, out of the 666 students scheduled for tutoring.

Fig. 9. Participation by Week



The participation rate is defined as the percentage of students who attended at least one tutoring session during the specified week. It is measured among students in schools that have started tutoring by that week.

Fig. 10. Dosage by Week



Dosage is defined as the average number of tutoring sessions per student in the specified week. It is measured among students who were scheduled for tutoring and who participated in at least one session during the monitoring period, and within schools that have started tutoring in that week.

How do dosage and participation vary by school?

Fig. 11. Tutoring Participation by School, between 07/31 and 09/01

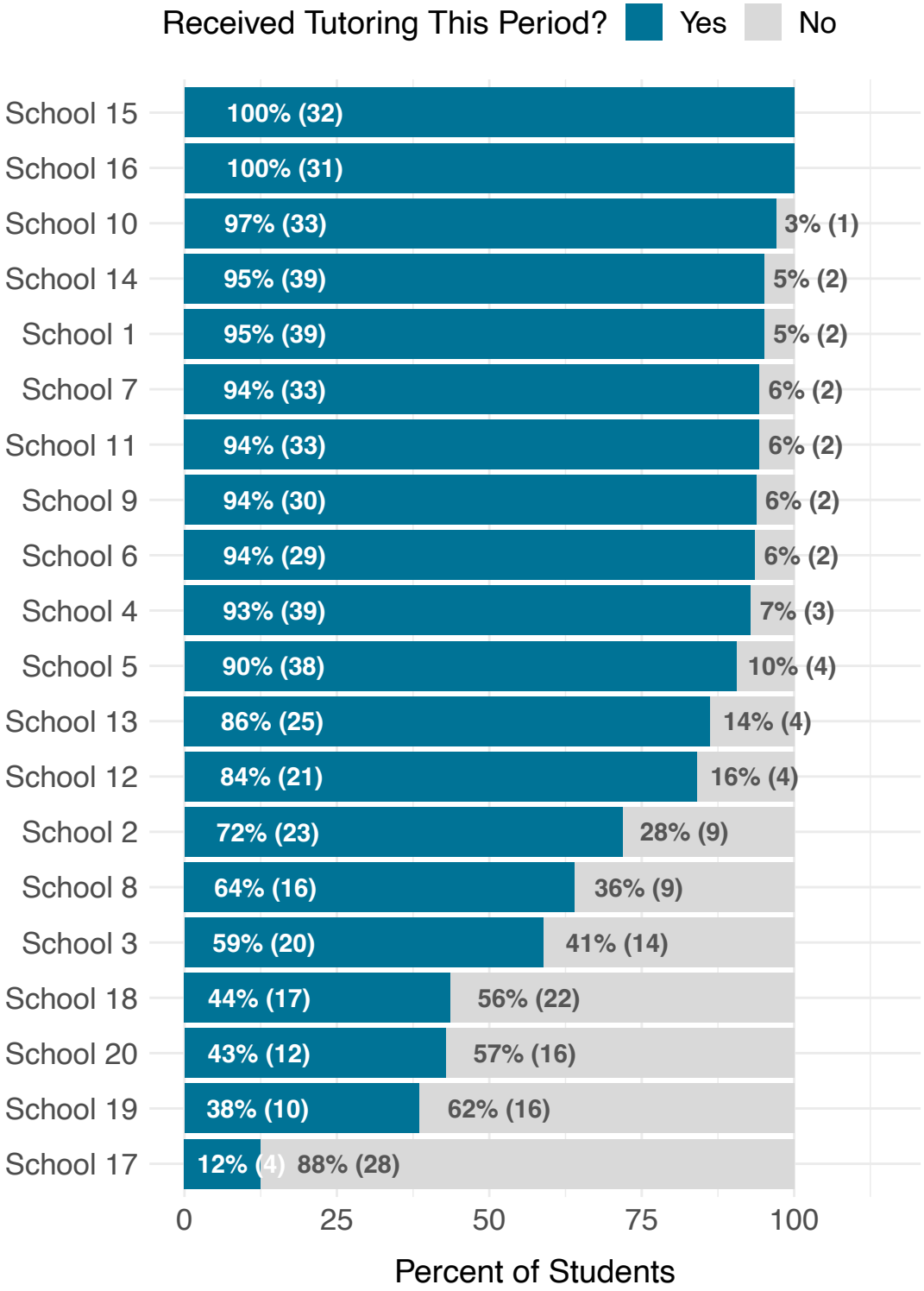


Fig. 12. Tutoring Dosage by School, between 07/31 and 09/01

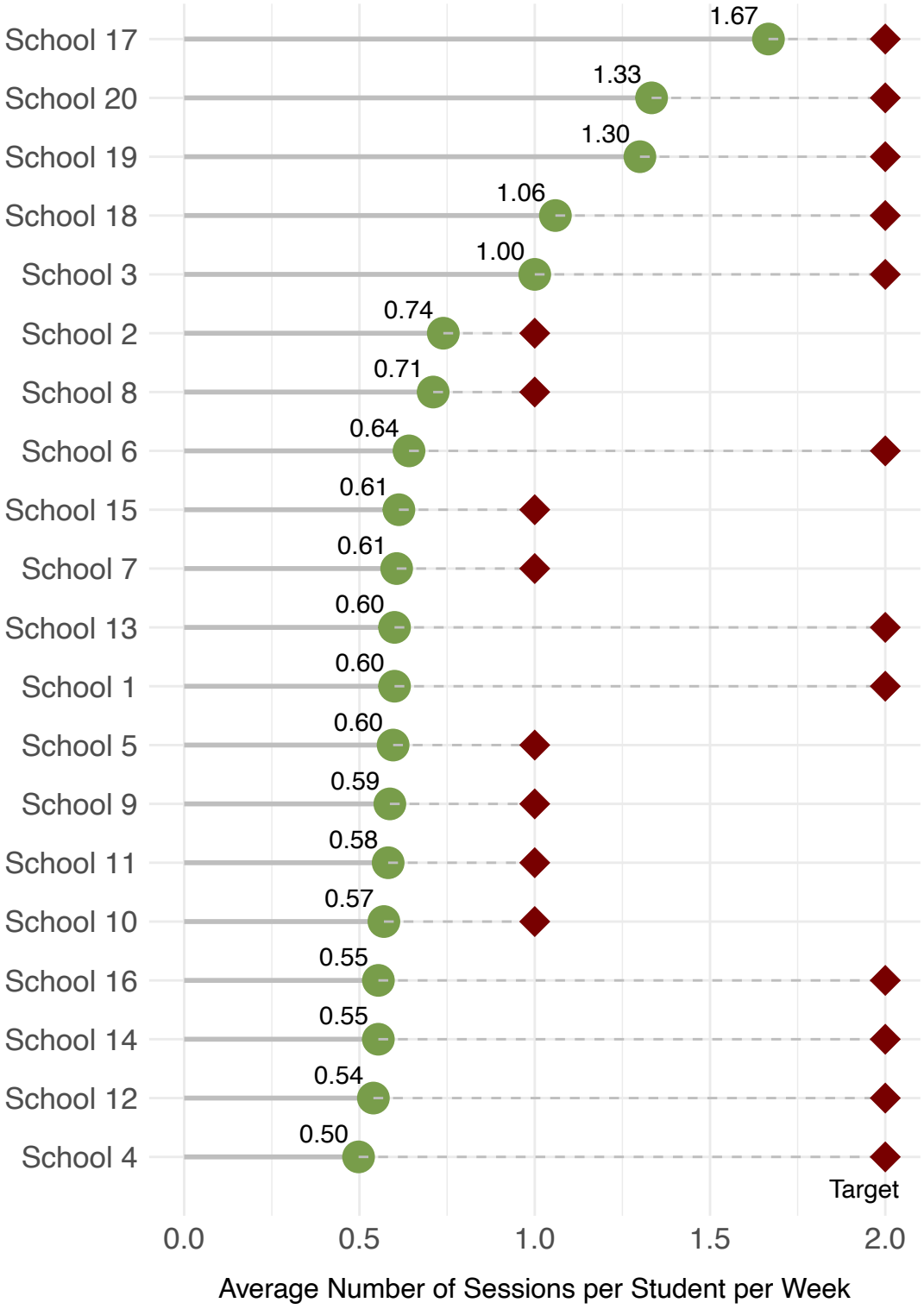
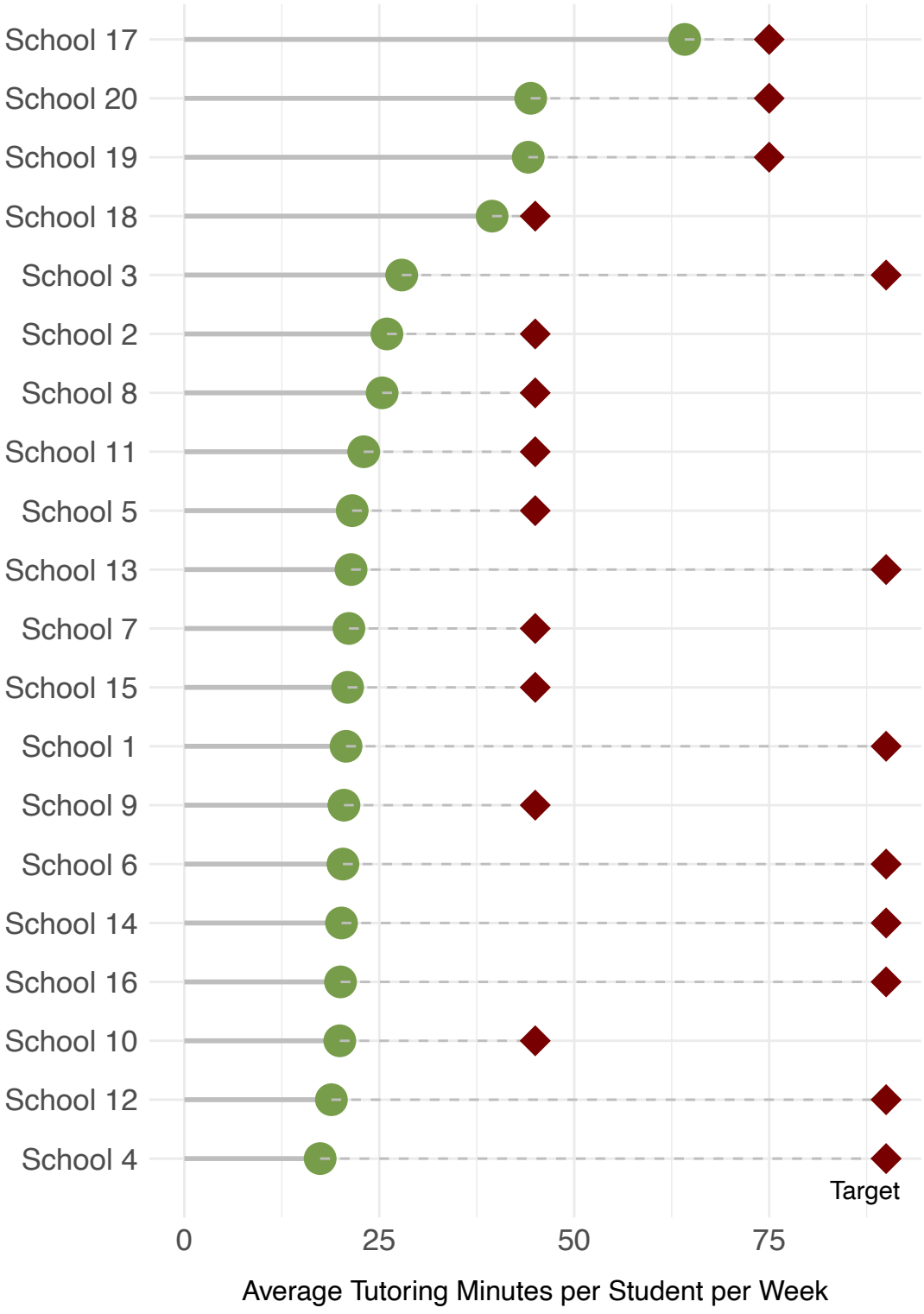


Fig. 13. Average Tutoring Minutes per Week by School, between 07/31 and 09/01



Supplementary Information

Data Description and Definitions

The focus of this report is on students in the impact study sample of the Personalized Learning Initiative. In PLI District, 20 schools have enrolled in the SY23-24 study. At those schools, 666 students were assigned to receive tutoring.

Data Sources

[Site teams will describe the participation data sources in the district]

Data Quality

Site teams will provide data quality concerns here, including but not limited to:
- Monitoring the extent of missing data over time
(missingness by schools/vendors/months)
- Interpretation of ambiguous attendance values
- How vacation days/weeks are handled

Definitions

Research Groups
High Dosage Tutoring (HDT): 231 students
Sustainable High Dosage Tutoring (SHDT): 435 students
Business As Usual (BAU): 334 students

Tutoring Start Date
Start dates for tutoring differ across schools. Specific dates for each school are provided at the end of this report.

Participation
Tutoring participation measures whether a student attended at least one tutoring session during the monitoring period. After an initial view of participation by research group (Table 1), we limit remaining participation statistics to students assigned to tutoring.

Dosage
Tutoring dosage measures how much tutoring students received during the monitoring period in terms of the number of sessions or minutes per week. Dosage statistics are limited to students who were assigned to receive tutoring AND who attended at least one session since the start of the study period.

Tutoring Start Dates

Table 6: Tutoring Start Date by School

School	Tutor Start Date
School 17	07/03/2023
School 8	07/04/2023
School 3	07/06/2023
School 15	07/06/2023
School 2	07/12/2023
School 6	07/12/2023
School 11	07/13/2023
School 9	07/17/2023
School 5	07/20/2023
School 16	07/24/2023
School 10	07/25/2023
School 14	07/25/2023
School 19	07/27/2023
School 4	07/28/2023
School 7	07/28/2023
School 1	07/31/2023
School 20	08/03/2023
School 18	08/04/2023
School 13	08/08/2023
School 12	08/14/2023

THE UNIVERSITY OF CHICAGO
EDUCATION LAB
Urban Labs

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY