Accelerate
The National Collaborative for Accelerated Learning

Contextualizing the Impact of Tutoring on Student Learning: Efficiency, Cost Effectiveness, and the Known Unknowns

RESEARCH REPORT

Prepared By
Luke Kohlmoos
Matthew P. Steinberg, Phd

May 2024

Acknowledgements

The authors thank the following individuals for their detailed and thoughtful feedback on multiple versions of this report: Kim Cassel (CasselView Advising); Gregory Chojnacki (Mathematica); Dan Goldhaber (CALDER at American Institutes for Research); Jeffrey Max (Mathematica); Meghan McCormick (Overdeck Family Foundation); Amanda Neitzel (CRRE at Johns Hopkins University); and Steve Ross (CRRE at Johns Hopkins University).

Matthew Steinberg, Managing Director of Research & Evaluation at Accelerate and corresponding author, may be contacted at: matthew.steinberg@accelerate.us.

About Accelerate

The National Collaborative for Accelerated Learning

Accelerate is a national nonprofit organization that serves as a central hub for proliferating effective, evidence-based academic interventions. By bridging gaps between research, policy, and school systems, Accelerate aims to embed proven high-dosage tutoring programs into public schools now and for the long term.

Accelerate funds innovation in schools, supports high-quality research, and advances a federal and state policy agenda to support this work. Accelerate was incubated and launched by America Achieves in 2022, and is a lead technical assistance partner to the National Partnership for Student Success (NPSS). In its initial years of work, Accelerate has made grants to over 60 programs across 29 states and has funded and supported over 65 research studies, including 21 randomized control trials.

Accelerate is supported by Citadel founder and CEO Kenneth C. Griffin; Arnold Ventures; the Bill & Melinda Gates Foundation; the Overdeck Family Foundation; and the Walton Family Foundation.

For more information, visit http://www.accelerate.us.

Accelerate

The National Collaborative for Accelerated Learning

Executive Summary

Existing evidence of tutoring program impacts indicates that high-dosage tutoring can significantly improve student learning, outperforming other educational interventions. Yet, tutoring is not a one-size-fits-all intervention. The design and implementation of tutoring varies widely across tutoring programs and service providers. It is therefore imperative to understand and contextualize the impact that different tutoring programs have on student learning. In this research report, we: (i) summarize existing evidence on tutoring program impact from Nickow et al. (2023), a recent meta-analysis of tutoring program impacts; (ii) construct a new measure of the efficiency of tutoring program impact, which we define as the hours of tutoring necessary to improve student learning by one month, and compare the efficiency of tutoring program impact across select tutoring providers based on evidence from well-designed randomized control trials (RCTs); (iii) propose an approach to measure the cost effectiveness of tutoring program impact, which we define as the additional months of student learning produced at a cost of $1,000 per pupil; and (iv) lay out a research agenda calling for significantly more evidence on the specific tutoring programs that improve student learning, for which students, in what educational contexts, at what fiscal cost, and the programmatic features most associated with student achievement gains.

We show that tutoring is an intervention with a long history of evidence in support of its effectiveness at improving student learning. However, the evidentiary base is less robust than initially expected. Only 15 of the 89 RCTs included in Nickow et al. (2023) would meet the highest standards of evidence on program impact. Our analysis finds that, among a select sample of tutoring providers, math tutoring more efficiently improves student learning (and with less variability) than literacy tutoring. We describe a research agenda which documents the need for more rigorous evaluations of tutoring program impact, especially for historically underserved student groups. We provide guidance to policymakers and researchers on the collection of valid programmatic cost data to calculate program-specific cost-effectiveness measures.

Our aim is for this report to provide policymakers, school leaders, and researchers with an approach to compare (and select) tutoring programs that most efficiently improve student learning at the lowest cost, and to support ongoing decision-making among educational leaders investing scarce resources to improve student learning.

Introduction

Tutoring not only produces significant gains in student learning but also outperforms other educational interventions (Kraft & Falken, 2021; Nickow et al., 2023). In a recently published peer-reviewed meta-analysis of 89 randomized control trials (RCTs) of tutoring interventions – meta-analysis is an approach for aggregating the impact of an intervention (such as tutoring) from across multiple studies – Nickow et al. (2023) show that tutoring improves student achievement by, on average, 0.29 standard deviations (SD), or approximately four additional months of learning for the typical elementary aged student. This meta-analytic effect, which was based on studies published between 1985 and 2020 of tutoring providers that served students in different grade levels and subject areas with different tutor types (e.g., teachers, paraprofessionals, parents), significantly outperforms other educational interventions, including class size reduction (0.13-0.20 SD); vacation academies (0.06-0.16 SD); summer school (0.08-0.09 SD); and extended school day/year (0.05 SD) (Kraft & Falken, 2021). Thus, tutoring is an educational intervention with a host of evidence to support not just its effectiveness for improving student achievement, but its relative dominance over other educational interventions. In a post-pandemic climate when schools across the country face an existential crisis of student achievement, tutoring is among the best solutions for accelerating learning recovery and closing persistent achievement gaps.[1]

Yet, tutoring is not a one-size-fits-all intervention. Tutoring as a generic term masks considerable heterogeneity in tutoring program design, the many different ways in which tutoring might be implemented in schools and classrooms, and in the effect sizes those program models have been shown to produce when rigorously evaluated (Nickow et al. 2023). This variability includes highly influential design decisions. Tutoring may occur during the school day or after school. Tutoring may be implemented in different classroom settings serving different student populations and grade levels, for varying lengths of time (i.e., dosage), and at different per pupil costs. Tutoring may also rely on different tutor types, including certified teachers, paraprofessionals, and community volunteers.

Informed by the extant literature, Accelerate defines high-dosage tutoring based on the following criteria: (i) tutoring is individualized, with a low student-to-tutor ratio (not higher than 4:1); (ii) tutoring is intensive, with 3-5 sessions/week for at least 30-60 minutes/session; (iii) tutoring is sustained, with a total of 50 hours of tutoring over 36 weeks, with some effective tutoring programs lasting for at least 10 weeks; (iv) tutoring is based on high-quality instructional materials (HQIM), and should incorporate direct instruction by tutors using HQIM and formative assessments aligned with grade-level content and Tier 1 classroom instruction - tutoring is not homework help; and (v) tutoring occurs during the school day since school-embedded programs have a significantly higher likelihood of student attendance than after-school programs (Kraft & Falken, 2021; Nickow et al., 2023; Robinson et al., 2021).

High-dosage tutoring has the potential, for example, to double the annual growth in high school math achievement (Guryan et al., 2023). Yet, the impact that tutoring programs and providers have on student learning varies widely (Nickow et al. 2023).

[1] Recent evidence indicates that most students in the United States will require a minimum of 3 school years to recover the academic learning that was lost following the pandemic (Kuhfeld & Lewis, 2022), and results from the 2022 administration of the National Assessment of Educational Progress (NAEP), our nation's academic report card, indicate that student math and reading achievement have regressed by two to three decades.

More research is needed to determine whether the impact of tutoring observed in prior research can be replicated in new school settings, for different student populations, under varying implementation conditions, and at greater scale. Attending to the nuances of tutoring program design and implementation is critical for identifying and selecting tutoring programs with the greatest likelihood of improving student outcomes in local settings and at scale.

Thus, it is essential to contextualize (the magnitude of) the impact that different tutoring programs - and programs with different design features - have on student learning. Doing so is necessary to provide policymakers and school leaders with critical information to compare and identify tutoring programs that most efficiently improve student learning at the lowest cost. A straightforward and policy-relevant approach to compare the efficiency and cost effectiveness of tutoring program impact across different providers is critical, particularly in a post-pandemic educational climate of student learning loss and in an increasingly constrained budgetary environment.[2]

Inside this research report, we:

1. Summarize existing evidence on tutoring program impact from the Nickow et al. (2023) meta-analysis;
2. Construct a measure of the efficiency of tutoring program impact, which we define as the hours of tutoring necessary to improve student learning by one month, and compare the efficiency of tutoring program impact across select tutoring providers based on evidence from well-designed RCTs;
3. Propose a measure of the cost effectiveness of tutoring program impact, which we define as the additional months of student learning produced at a fixed (per pupil) cost of $1,000; and
4. Identify a research agenda documenting what is still unknown about the design, implementation, and impact of tutoring.

This research report leverages existing evidence on tutoring program impact to provide guidance to policymakers and school leaders on the tutoring providers that most efficiently improve student learning. As additional evidence on tutoring program impact from well-designed RCTs becomes available, including from Accelerate-sponsored evaluations and evaluations conducted in the field more broadly, we will expand our dataset and update evidence to inform comparisons across providers on the efficiency of tutoring program impact. Ultimately, our goal is to support ongoing decision-making among educational leaders investing scarce resources in tutoring as a key driver for improving student learning.

[2] In response to the pandemic, the federal government provided $190 billion of stimulus aid to the K-12 public education system via the Elementary and Secondary School Emergency Relief Fund (ESSER). ESSER funds officially expire in September 2024. According to the Education Recovery Scorecard, as of January 2024, $51 billion in ESSER funds remain unspent or unallocated.

What We Know: Findings from Nickow et. al (2023)

Nickow et al. (2023) show that tutoring has a significant and substantively positive impact on student achievement across a wide range of tutoring programs, on the order of 0.29 SD.[3] Yet, a closer assessment of the specific programmatic and research design features across multiple experimental evaluations of tutoring programs provides a more nuanced understanding of the impact of tutoring – including the existence of well-designed and well-implemented program models that have been found to produce no impact on student achievement – and the existing gaps in the current evidentiary base on tutoring as an educational intervention.

Table 1 summarizes the evidence from Nickow et al. (2023) by select program and research design features. On average, among the 89 RCTs included in the meta-analysis, small-scale studies – those with smaller sample sizes – tend to have greater effects on student learning than large-scale studies.[4] Tutoring programs with teacher-led tutoring have, on average, larger effects than tutoring provided by non-professionals and parents. Tutoring providers that offer more intensive dosage (at least three sessions per week of tutoring) have greater impacts on student learning than providers with less intensive dosage (fewer than three sessions per week).

It is especially important to underscore the evidence showing that the magnitude of tutoring program impact varies among experimental evaluations with larger and smaller sample sizes. A primary goal of tutoring as a pandemic recovery strategy is to reach significantly more than just the 10 percent of students nationally who currently receive high-dosage tutoring. So, the extent to which the impacts of tutoring might be scaled to reach more students is informed by the impact of tutoring in studies with large sample sizes, of which there are few large scale impact evaluations on tutoring (Nickow et al., 2023). Additionally, scaling up smaller scale programs is challenging, both in terms of maintaining high standards of implementation fidelity but also in terms of the logistics necessary to hire and train more tutors, develop productive relationships with more school sites, and collect and analyze more data on implementation and impact.

Notably, while program-specific differences in tutoring impact provide insight into potential differences across program design features, it is important to avoid drawing causal conclusions

[3] The share (or proportion) of the average annual growth in student achievement (i.e., the additional months of learning) that corresponds to a standardized effect size of 0.29 SD depends on the grade level and subject area. For example, a 0.29 SD effect size corresponds to 3 additional months of math growth for grade 3 students and 12 additional months of math growth for grade 9 students; a 0.29 SD effect size corresponds to 4 additional months of reading growth for grade 3 students and 11 additional months of reading growth for grade 9 students. As described in Hill et al. (2008), the estimated average annual growth in student achievement (upon which these month-specific estimates of learning gains are based) are measured from spring to spring and capture changes in student learning across a year's time, including learning in school, learning and maturation experienced outside of a school setting, and any changes in student learning (such as summer melt) occurring during the summer (non-school) months. See Hill et al. (2008) and Kraft (2020) for a more detailed discussion of the variability in average annual achievement growth across grades and subjects.

[4] Among small-scale studies, the average impact of tutoring was 0.44 SD in 18 program evaluations with a study sample of less than 51 students, and 0.36 SD in 21 program evaluations with a study sample of 51-100 students. In contrast, among larger-scale studies, the average impact of tutoring was 0.18 SD in 12 program evaluations with a study sample of 201-400 students, and 0.20 SD in 15 program evaluations with a study sample of more than 400 students. This is a widely-observed phenomenon in social policy research, and may reflect, for example, greater researcher and/or program provider contributions to service delivery relative to typical implementation conditions; reliance on shorter-term or more proximal measures of achievement (e.g., early literacy skills as opposed to more summative measures of reading comprehension); and/or reliance on researcher- or provider-developed measures of student achievement which may be more closely aligned to the tutoring content.

from these program design trends described in Table 1. These program features were not tested experimentally - that is, program features such as tutor type or dosage were not randomly assigned across students. As a result, differences in average impacts that are linked with these features might be driven by other factors. And while tutoring was randomly assigned across students, there are likely to be interactions between different program design features and student selection. For example, early grade tutoring programs are likely to have shorter sessions than older grades. This means that the larger impacts observed among programs with shorter sessions might actually be due to younger students who benefit more from each tutoring minute than older students (i.e., the returns to dosage might vary across the grade and age distribution). Teacher-led tutoring is more likely to be conducted during the school day than tutoring sessions conducted by parents or other nonprofessionals who are not certified teachers.

Table 1. Summary of Findings in Nickow et al. (2023), by Program/Study Design Feature

Program or Study Feature	Key Findings
Study Sample	• Of the 89 studies, 15 had a study sample of at least 350 students • Pooled effect of tutoring: 0.29 SD • Small scale studies: (i) less than 51 students: 0.44 SD; (ii) 51-100 students: 0.36 SD • Large scale studies: (i) 201-400 students: 0.18 SD; (ii) more than 400 students: 0.20 SD • Studies published 1985-2009: 0.35 SD; studies published after 2010: 0.25 SD
Subject	• Literacy outcomes account for 81% of study sample • Literacy (0.29 SD) and math (0.27 SD) have similar overall effects • Literacy has more variability in program effects than math
Grade	• 10 (of 89) studies include grades 6-11 • Effect sizes decline as grade level increases • All grade levels have effect sizes ~25-30% of expected growth (when normalized by each grade level's average annual growth) • Grades 6-11 math have effect sizes of 71% of expected growth, but across just 5 studies
Tutor Type	• Teacher-led tutoring (19 studies): 0.39 SD • Paraprofessional-led tutoring (43 studies): 0.30 SD • Non-professional-led tutoring (21 studies): 0.17 SD • Parent-led tutoring (9 studies): 0.23 SD
Dosage	• <3 times/week: 0.10 SD • 3 times/week: 0.29 SD • 4-5 times/week: 0.35 SD • Sessions less than 30 minutes outperform those that are longer than 30 minutes, but this may be due to heavy weighting towards K and 1 programs • Programs that are 1 semester or less (0.30 SD) tend to outperform longer programs (0.26 SD), but the pattern is unclear by grade level
Tutor-Student Ratio	• 1:1 Ratio: 0.32 SD (60 studies) • 1:2 Ratio: 0.26 SD (15 studies) • 1:3+ Ratio: 0.26 SD (23 studies)

Thus, while instructive about the potential differences in student achievement associated with different program design features, the Nickow et al. (2023) meta-analysis (as with other syntheses of existing studies) provides little guidance as to which program design features may (or may not) generate improvements in student learning. Understanding how specific program design features of tutoring programs impact student learning is an area in need of significantly more research, which we describe in the context of a broader research agenda at the end of this report.

And while Nickow et al. (2023) report on program impacts in standardized effect sizes (i.e., standard deviation units), effect sizes of program impact alone tell us little about the relative impact of tutoring across program providers (i.e., tutoring efficiency) and nothing about the cost effectiveness of tutoring. Finally, even though Nickow et al. (2023) included 89 RCTs in their meta-analysis, just 15 of those RCTs would meet ESSA Tier 1 evidence standards based on sample size criteria - namely, that the study include a minimum of 350 study participants.

Section Two

Our Approach

Below, we first describe our sample selection criteria for identifying and constructing an analytic sample of RCTs of tutoring program impact. Next, we describe the criteria for selecting the measures of student achievement (i.e., outcome measures) included in our analysis. We then describe how we construct a measure of the efficiency of tutoring program impact, and conclude by proposing a model for documenting the programmatic costs associated with tutoring providers and calculating the cost effectiveness of tutoring impacts.

Sample Selection Criteria for RCTs of Tutoring Program Impact

We first identified RCTs from Nickow et al. (2023) with at least 350 study participants. We include only studies with sufficient sample size to meet ESSA Tier 1 evidence standards.[5] This sample selection criteria yielded 15 studies. Of these 15 studies from Nickow et al. (2023), we excluded five studies that were either conducted in an international setting, not focused on grades K-12, or provided homework help (instead of tutoring). Two studies were conducted in Ireland; a third study did not report the amount of time that students received tutoring each week (i.e. dosage); a fourth study was in pre-K; and a fifth study focused on a program that provided homework help and was absent a specific subject or curriculum.

Outreach to colleagues in the tutoring field familiar with recent evidence on program impact led to the identification of four additional studies that were not included in Nickow et al. (2023) but that both met ESSA Tier 1 evidence standards for sample size (i.e., the study included a minimum of 350 study participants) and were published in a peer-reviewed journal. These included two studies of Future Forward (Jones et al., 2018 and 2021); one study of a New York City ELA program (Fryer et al.,

[5] A minimum sample of 350 students is sufficient to power a study to detect impacts on student achievement consistent with what might be expected from a tutoring program (0.15-0.20 SD). Even larger sample sizes are necessary to detect differences in program impact by student characteristics, and to produce more precise estimates of the average treatment effect. An additional benefit of a study sample of at least 350 students is it requires the program being evaluated to operate at sufficient scale.

2020); and one study replicating prior research on Minnesota Reading Corps (Markovitz et al., 2022).[6]

In the Reference section, we document all studies that were considered for and ultimately included in our analytic sample, which we denote as the following: (i) studies included in the re-analysis analytic sample that were included in Nickow et al. (2023); (ii) studies included in the re-analysis analytic sample that were not included in Nickow et al. (2023); (iii) studies that were included in Nickow et al. (2023) with a study sample greater than 350 but not included in the re-analysis analytic sample because they did not meet all selection criteria for the re-analysis; and (iv) studies that we reviewed that were not included in Nickow et al. (2023) and which were not included in the re-analysis analytic sample because they did not meet all selection criteria for the re-analysis.

Selection Criteria for Outcome Measures

Once we identified the sample of RCTs of tutoring program impact to include in our analytic sample (which include 14 RCTs of 12 tutoring programs), we next selected the most relevant outcomes from each of these 14 RCT studies.[7] We focused on the most policy-relevant measures of math or reading performance available in the study, prioritizing measures from assessments of academic performance that are in widespread use. These are likely the assessments most generalizable to state assessments and most commonly used for decision-making within schools. Typically each study included 1-2 outcome measures from standardized assessments such as DIBELS, Aimsweb or SAT-10. Notably, only one of the included studies relied on end of year state assessments, although some assessments have been used in state accountability systems (SAT-10, Iowa Test of Basic Skills, ACT's EXPLORE and PLAN). We identified a total of 18 outcome measures across the 14 RCTs, 17 of which showed statistically significant tutoring program effects.

The included outcomes were contemporaneous to the year in which the tutoring occurred. We note that this is a key limitation of the underlying studies, as the absence of longer-term outcomes limits insight into whether (and the extent to which) the impact of tutoring persists (or fades away) over time. However, only two studies included in our analytic sample had outcome measures assessing student achievement a semester or more after the conclusion of tutoring (Clarke et al., 2017 and Fryer et al., 2020). A third study focused solely on longer term outcomes for a program included in the analytic sample and found positive results (Jones et al., 2023) but was excluded because it did not include any outcomes from the year in which the tutoring took place. The inclusion of longer-term outcomes in evaluations of tutoring impact is an important area of future research since there is limited evidence on the persistence of tutoring impacts beyond the school-year in which tutoring has been provided (which, as noted below, has important implications for assessing the cost effectiveness of tutoring).

Finally, we included Intent-to-Treat (ITT) estimates of program impact and excluded Treatment-on-the-Treated (TT) or Optimal Dosage estimates. Arguably, ITT estimates are the most policy relevant estimate of program impact since they provide policymakers with evidence on the impact that could be expected if a program is adopted (i.e., ITT estimates provide evidence to inform the policy question: what would have happened to student achievement if tutoring was offered to eligible

[6] Of the 14 studies included in our analytic sample, 10 were reviewed by What Works Clearinghouse (WWC), of which 8 met WWC standards without reservation.

[7] Outcomes that were excluded tended to be narrower in focus (i.e., discrete or specific skills) and/or researcher- or tutoring provider-created assessments. Examples include curriculum-based measures of oral counting, sight word assessments, and letter identification subtests for students older than kindergarten.

students?).[8] The one exception was the inclusion of TT estimates from the two studies of Saga included in Guryan et al. (2023) since the randomization of tutoring in that study occurred prior to the start of the school year and there was substantial post-randomization attrition from the study among students who didn't end up attending the expected school.

Together, the sample selection criteria for the RCTs of tutoring program impact and associated outcome measures yielded a study sample of 14 RCTs of 12 tutoring programs covering 18 academic outcomes that were published in peer-reviewed journals. All 14 RCTs have been published since 2012. Table 2 summarizes the 12 tutoring programs included in the analytic sample, as well as each tutoring program's design features, including the subject and grade levels served by the tutoring program, the tutor type, student-tutor ratio, and the outcome measures associated with each tutoring program. We also indicate (where available) the counterfactual condition (i.e., the services that control group students received in lieu of receiving tutoring as part of the treatment group). Across the full analytic sample of 12 tutoring providers and 18 outcomes, eight outcomes (and thus eight impact estimates) are focused on reading/literacy and ten outcomes (and associated impact estimates) are focused on math. Of the 12 tutoring providers, nine focus on elementary school grades (i.e., K-5), two focus on middle school grades (i.e., 6-8), and one focuses on high school grades (i.e., 9-12). Further, eight (of 12 tutoring providers) rely on paraprofessionals to deliver tutoring, two rely on certified teachers, and two rely on nonprofessionals.[9] Interestingly, the fact that most providers relied on paraprofessionals to provide tutoring suggests the popularity of supplementing the work teachers do with AmeriCorps fellows, part-time university students and other adults in the building such as interventionists. The student-tutor ratio in which tutoring is delivered among our sample of tutoring providers tends to be on a one-to-one or two-to-one basis. The counterfactual condition varies across the selected studies, with business-as-usual approaches sometimes including access to other supplementary/intervention services.

[8] There are several reasons why Intent-to-treat (ITT) and treatment-on-the-treated (TT) estimates may differ in policy-relevant ways. Some programs may be highly effective with treated students, but due to implementation challenges, few students may actually receive the intended treatment. Indeed, students who might benefit the most from tutoring may be the same students most likely to be eligible for tutoring but who do not actually receive tutoring. A study of academic recovery strategies in 12 mid- to large-sized school districts across 10 states during the 2021–22 school year found that, on average, only 20-30 percent of eligible students received high-dosage tutoring corresponding to just 5-10 percent of all students in schools with high-dosage tutoring (Carbonari et al., 2022). And among those eligible students receiving high-dosage tutoring, students received just 12-14 hours of math tutoring, on average, in districts that planned to offer 30-60 hours of math tutoring, and the average dosage of tutoring received by students ranged from 4-10 hours (Carbonari et al., 2022). Thus, the implementation conditions and contextual features in which tutoring is provided are key to understanding program impacts. A more nuanced and empirical assessment of the contextual factors that might be associated with program impacts, and the conditions that might support (or constrain) implementation fidelity, are key to understanding how tutoring might be implemented at scale.

[9] Following Nickow et al. (2023), we focus on four categories of tutor types: teachers; paraprofessionals; nonprofessionals; and parents. Teachers are certified teachers; paraprofessionals may include educators who are not certified teachers, including school staff members, education students, and professional development fellows such as AmeriCorps; nonprofessionals are volunteer tutors, such as community residents and retired adults. Notably, Nickow et al. (2023) exclude tutoring interventions led by peer tutors.

Notes about Table 2. Following Nickow et al. (2023), we define tutor type by four categories: Teachers, in which certified teachers serve as tutors; Paraprofessionals, in which non-certified teachers are employed as tutors such as education students, Americorps Fellows or other school-based staff; Nonprofessional, or volunteer in which tutors are not professionally engaged as tutors; and Parents or caregivers in which programs support parents or caregivers to tutor their children. The analytic sample did not include any parent/caregiver programs. Future Forward Impact is Jones et al. (2021) using only the year one outcome. Future Forward Pilot is Jones et al. (2018) using the outcomes following the second year of the treatment. Reading Recovery is Sirinides et al. (2018). Experience Corps is Lee et al. (2020). Minnesota Reading Corps Replication is Markovitz et al. (2020) using only grades 2-3 Minnesota outcome that met the sample size criteria. Reading Partners 1 and 2 are both Jacob et al (2016). NYC ELA is Fryer et al. (2020) using only the year one outcome. Roots is Clarke et al. (2017). High 5s 1 and 2 are both outcomes from Mattera et al. (2018). Number Rockets/Galaxy Math 1 and 2 are two different treatments from Fuchs et al. (2013). Number Rockets/Galaxy Math 3 is a replication study by Gersten et al. (2015). Math Recovery is Smith et al. (2013). Minnesota Math Corps is Parket et al. (2019). Saga 1 and Saga 2 are two studies that were combined in Guryan et al. (2023).

Table 2. Summary of Tutoring Programs, By Program Design and Outcome Measures

Tutoring Program	Subject	Grade	Tutor Type	Student:Tutor Ratio	Outcome Metric	Counterfactual Condition
Future Forward Pilot	Literacy	K-3	Paraprofessional	1 : 1	MAP	Business as Usual
Future Forward Impact	Literacy	K-3	Paraprofessional	1 : 1	MAP	Business as Usual (standard literacy instruction, additional support programs unknown)
Reading Recovery	Literacy	1	Teacher	1 : 1	Iowa Test of Basic Skills - Total Reading	Business as Usual (access to normally provided literacy supports)
Experience Corps	Literacy	1-3	Nonprofessional	1-3 : 1	Woodcock Johnson Passage Comprehension	Business as Usual
Minnesota Reading Corps Replication	Literacy	2-3	Paraprofessional	1 : 1	Aimsweb - Oral Reading Fluency	Business as Usual
Reading Partners 1	Literacy	2-5	Nonprofessional	1 : 1	SAT-10	Business as Usual (eligible for other supplemental services)
Reading Partners 2	Literacy	2-5	Nonprofessional	1:1	Aimsweb - Oral Reading Fluency	Business as Usual (eligible for other supplemental services)
NYC ELA	Literacy	6	Paraprofessional	4:1	State test score	Business as Usual
ROOTS	Math	K	Paraprofessional	2:1 or 5:1	TEMA	Business as Usual
High 5s 1	Math	K	Paraprofessional	3-4:1	REMA-K	Business as Usual (80% whole group with single teacher)
High 5s 2	Math	K	Paraprofessional	3-4:1	Woodcock-Johnson Applied Problems	Business as Usual (80% whole group with single teacher)
Number Rockets / Galaxy Math 1	Math	1	Paraprofessional	1:1	Word Problems	Business as Usual
Number Rockets / Galaxy Math 2	Math	1	Paraprofessional	1:1	Word Problems	Business as Usual
Number Rockets / Galaxy Math 3	Math	1	Paraprofessional	2-3:1	TEMA-3	Business as Usual (No additional supplemental math services)
Math Recovery	Math	1	Teacher	1:1	Woodcock Johnson III - Math Reasoning (AP+QC)	Business as Usual
Minnesota Math Corps	Math	4-8	Paraprofessional	2:1	STAR Math	Business as Usual (eligible for other supplemental services)
Saga 1	Math	9-10	Paraprofessional	2:1	EXPLORE/PLAN	Typically "double dose" algebra
Saga 2	Math	9-10	Paraprofessional	2:1	EXPLORE/PLAN	Typically elective course like PE or Art

Measuring the Efficiency of Tutoring

To enable a more apples-to-apples comparison of the impact of tutoring across tutoring program providers, we construct a measure of the efficiency of tutoring program impact. The efficiency of tutoring program impact is measured as the hours of tutoring necessary to improve student learning by one month. We rely on the experimental estimates of tutoring program impact included in our analytic sample to calculate tutoring efficiency. In Table 3, we present 18 experimental estimates (Effect Size) across the 12 tutoring programs included in our analytic sample.

To construct our measure of tutoring efficiency, we first normalize the program-by-outcome effect (Effect Size) by grade level and subject because relative gains in student achievement vary by grade and subject. This normalization is necessary because, for example, an effect size of 0.15 SD represents a larger gain in student learning for students in grade 6 than for students in grade 1 because the average annual growth in student achievement (in standard deviation units) is larger for lower grades than for higher grades (Hill et al., 2008). To do so, we divide the program-by-outcome effect size by the average annual growth in student achievement at the grade and subject levels (Hill et al., 2008), which we denote as Average Annual Growth (SD). This yields the Percent of Average Annual Growth in student achievement generated by a specific tutoring program. Next, we calculate the number of months of additional learning growth generated by a tutoring program. To do so, we first multiply the Percent of Average Annual Growth by 180 (the total number of school days in an academic year); then, we divide this result by 20, or the total number of days of school in a month. This yields the number of additional months of student learning generated by a tutoring program.

Finally, we calculate **Tutoring Efficiency** - the hours of tutoring required to gain an additional month of student learning. To do so, we divide the Intended Dosage (in hours) of a tutoring program (i.e., the amount of tutoring a student receives as designed for the relevant tutoring program and study setting) by the number of additional months of student learning generated by a tutoring program.[10] This yields our measure of Tutoring Efficiency, measured as the number of hours of tutoring needed for one month of student learning. This measure of program efficiency enables a comparison of the impact of tutoring across programs implemented in different grades with different levels of program dosage. For example, increasing student learning by one month in 20 hours is more efficient than requiring 50 hours to increase student learning by one month.[11]

Tutoring Efficiency: Hours of tutoring to improve student learning by one month

$$\text{Tutoring Efficiency} = \frac{\text{Hours of intended dosage}}{\text{\# of additional months of learning gained by tutoring}}$$

[10] We rely on intended dosage because (i) it is more consistently reported than actual dosage; and (ii) it more accurately represents the amount of effort required to implement the program via changes in school schedules, the amount of tutor time required to implement with fidelity, and the number of students who can be served within existing time constraints.

[11] Note that scaling tutoring efficiency requires an assumption of linearity. For example, if a program requires 20 hours of tutoring to increase student learning by one month, the assumption of linearity assumes that 40 hours of tutoring would produce two additional months of student learning. Further investigation is required to better understand the extent to which linearity is, or is not, a valid assumption for the returns to tutoring time as captured by the tutoring efficiency measure.

Comparing the Efficiency of Tutoring Program Impact Across Providers

Table 3 summarizes the program impact estimates and our measure of tutoring efficiency across the 12 tutoring providers (and 18 associated impact estimates). In Table 3, we report the following key data points for each program outcome:

- effect size;
- average annual growth in student learning;
- percent of average annual growth in student learning for the relevant tutoring program impact estimate;
- the additional months of learning corresponding to the tutoring program impact estimate;
- the intended dosage (in hours) of tutoring provided by a tutoring program; and
- our measure of tutoring efficiency, measured as the hours of tutoring needed for an additional month of student learning.

Table 3 not only shows how the average annual growth - and the corresponding percent of average annual growth - depends on the grade and subject in which tutoring takes place, but also shows how accounting for this variability enables a more appropriate comparison of the impact of tutoring programs. For example, Saga Study 1 reports an effect size of 0.18 SD, similar to four other programs that had effect sizes within 0.05 SD of 0.18 SD (Experience Corps, High 5s REMA-K, Number Rockets/Galaxy Math Speeded Practice, Minnesota Math Corps). Yet, once we account for variation in the grade-by-subject in which tutoring occurs, Saga Study 1's 0.18 SD effect size corresponds to 6.9 months of additional learning, outpacing all other tutoring programs in our analytic sample other than the Saga Study 2. The importance of tutoring dosage is also apparent. For example, ROOTS delivers two months of additional learning but with a level of dosage (16.7 hours) lower than every other program included in the sample, and thus requires just 8.4 hours of tutoring to produce an additional month of student learning, more efficient than the average math tutoring program.

Figure 1 shows the distribution of tutoring efficiency by subject and tutoring program. In Figure 1, we find that, among our analytic sample, math tutoring is more efficient than literacy tutoring. Specifically, we find that tutoring providers focused on math instruction require, on average, 13.7 hours of tutoring to increase math achievement by one month; in comparison, providers focused on literacy instruction require, on average, 39.6 hours of tutoring to increase literacy achievement by one month.[12] There is also less variability in tutoring efficiency among math programs (SD=6.9 hours) than literacy programs (SD=25.3 hours). Notably, since there are few tutoring programs that met the sample selection criteria for inclusion in our analytic sample, the mean and variance of tutoring program efficiency (by subject) will be sensitive to the inclusion of additional studies and program impact estimates. This point further underscores the dearth of rigorous, large-sample experimental evaluations of tutoring program impact currently in the extant literature. Thus, additional high-quality RCTs of tutoring programs are required to provide greater insight into whether the relatively more efficient returns to math versus literacy tutoring observed in our analytic sample generalizes to a larger sample of tutoring program impacts across a wider variety of tutoring programs.

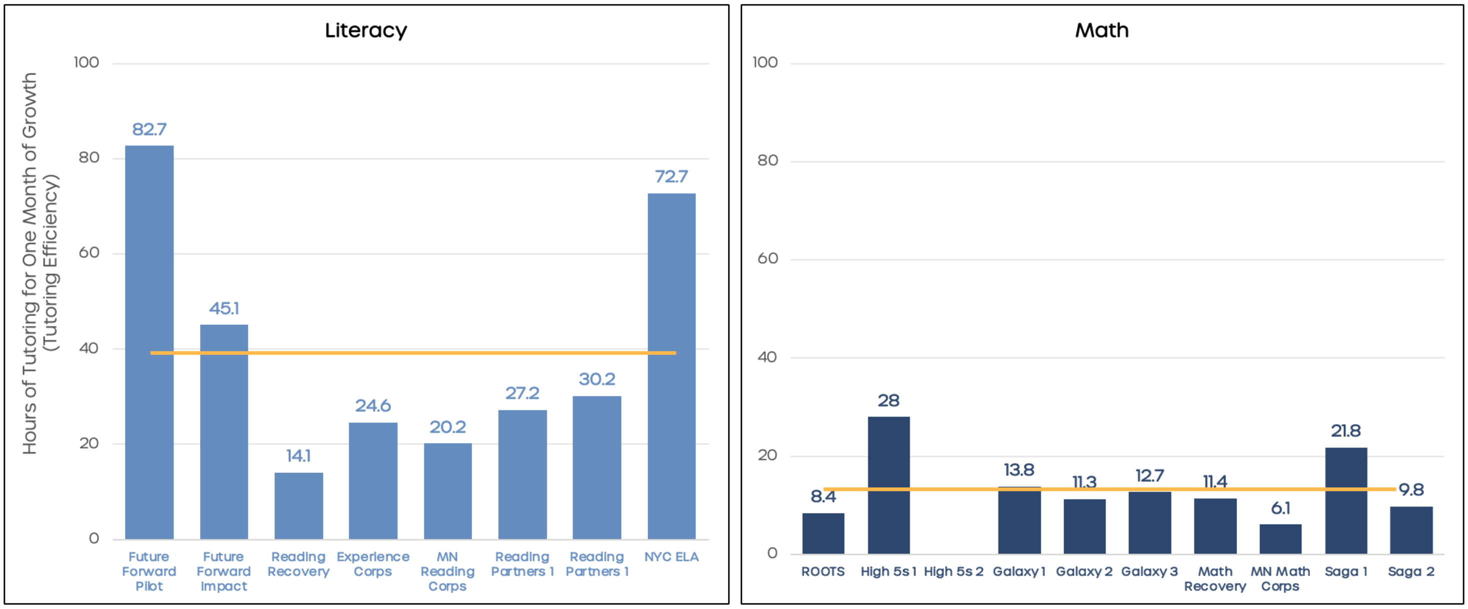
[12] This finding is consistent with a broader literature on the subject-specific impacts of educational interventions. Schooling effects tend to be larger in math than literacy, and programs and policies that improve student outcomes are more likely to impact math than literacy (Rich, 2013).

Table 3. Summary of Tutoring Programs, By Effect Size and Tutoring Efficiency

Tutoring Program	**Effect Size** (SD)	**Avg. Annual Growth** (SD)	**% of Avg. Annual Growth**	**Months of Additional Learning**	**Intended Dosage** (hours)	**Tutoring Efficiency** (hours of tutoring needed for one month of student learning)
Future Forward Pilot	0.12	1.153	10%	0.9	77.5	82.7
Future Forward Impact	0.11	1.153	10%	0.9	38.75	45.1
Reading Recovery	0.48	1.52	32%	2.8	40	14.1
Experience Corps	0.13	1.03	13%	1.1	28	24.6
Minnesota Reading Corps Replication	0.23	0.785	29%	2.6	53.3	20.2
Reading Partners 1	0.10	0.5825	17%	1.5	42	27.2
Reading Partners 2	0.09	0.5825	15%	1.4	42	30.2
NYC ELA	0.08	0.32	23%	2.1	153.3	72.7
ROOTS	0.25	1.14	22%	2.0	16.7	8.4
High 5s 1	0.19	1.14	17%	1.5	42	28.0
High 5s 2	0.0	1.14	0%	0.0	42	N/A
Number Rockets / Galaxy Math 1	0.22	1.14	19%	1.7	24	13.8
Number Rockets / Galaxy Math 2	0.27	1.14	24%	2.1	24	11.3
Math Recovery	0.30	1.14	26%	2.4	27	11.4
Minnesota Math Corps	0.17	0.422	40%	3.6	22	6.1
Saga 1	0.18	0.235	77%	6.9	150	21.8
Saga 2	0.40	0.235	170%	15.3	150	9.8
Mean (SD)	**0.20 (0.12)**	**n/a**	**32% (38%)**	**2.9 (3.4)**	**55.9 (45.9)**	**25.9 (21.9)**

Notes. Effect Size is the impact estimate for the selected outcome for the tutoring program. Average Annual Growth based on estimates from Hill, et al (2008) for each grade and subject. Percent of Average Annual Growth is calculated as the Effect Size divided by Average Annual Growth. Months of Additional Learning is the Percent of Average Annual Growth times 180 (to give days of growth out of a 180 day school year) divided by 20 (to give months of growth assuming 20 days of school per month). Intended Dosage is the designed dosage from the relevant tutoring program (and study setting). Intended Dosage retrieved from each study when session length, session frequency and duration of program were provided. In cases where a range of values was provided for session length, frequency or duration, the midpoint of the range was used to calculate intended dosage. In the case of Minnesota Math Corps, duration of the program varied across sites; we used the average duration (14.68 weeks) to calculate dosage (although at least 12 weeks was defined as optimal dosage). Most studies had high quality implementation in which it is either reported or implied that actual dosage approached intended dosage. Hours of Tutoring Needed for a Month of Additional Growth was calculated by converting Percent of Average Annual Growth to total months of growth, assuming 180 days of school per year and 20 days of school per month, and dividing Intended dosage by total months of growth. For the calculation of sample mean (SD) for Tutoring Efficiency, we excluded High 5s 2 since it has a value of undefined (i.e., N/A) on the tutoring efficiency measure because the impact estimate for High 5s 2 was not statistically significantly different from zero.

Figure 1. Distribution of Math and Literacy Efficiency Estimates, By Tutoring Program



Notes. Mean for Literacy is 39.6 hours (standard deviation is 25.3); mean for math is 13.7 hours (standard deviation is 6.9). High 5s 2 has been excluded from mean calculations because a program with no effects requires infinite hours to achieve an additional month of learning. Future Forward Impact is a K-3 literacy program using MAP as an outcome. The program is two years, the included outcome is the one year impact, selected due to high attrition rates from year one to year two. Future Forward Pilot is a K-3 Literacy Program using MAP as an outcome. It is a two year program and this outcome uses the two year impact and dosage. Reading Recovery is a grade 1 program using Iowa Test of Basic Skills as an outcome. Experience Corps is a grades 1-3 literacy program using Woodcock-Johnson Passage Comprehension as an outcome. Minnesota Reading Corps Replication is a K-3 literacy program. The selected outcome is Aimsweb Oral Reading Fluency andpools grades 2 and 3 because it is the only randomization grouping that met the minimum sample criteria. Reading Partners 1 is a grade 2-5 program using SAT-10 as an outcome. Reading Partners 2 is a grade 2-5 program using Aimsweb Oral Reading Fluency as an outcome. The two Reading Partners outcomes come from the same study. The New York City ELA program uses state assessments as an outcome. The included outcome focuses on grade 6, one year impacts, as future years of the study had high rates of attrition and grade 6 was most comparable to other included studies and outcomes. Roots is a Kindergarten program using TEMA as an outcome. High 5s 1 is using REMA-K as an outcome and is a Kindergarten program. High 5s 2 is using Woodcock-Johnson Applied Problems as an outcome and is a Kindergarten program. The High 5s 1 and 2 outcomes are generated from the same study. Number Rockets 1 is for the speeded practice program design and is a grade 1 program using a Word Problems assessment as an outcome. Number Rockets 2 is for the non-speeded practice program design and is a grade 1 program using a Word Problems assessment as an outcome. Number Rockets 1 and 2 outcomes are from the same study. Number Rockets 3 is a replication study and is a grade 1 program using TEMA as an outcome. Math Recovery is a grade 1 program using Woodcock-Johnson Math Reasoning as an outcome. Minnesota Math Corps is a grades 4-8 program using STAR Math as an outcome. Saga 1 is from the 2013-14 school year and only included boys and is a grades 9-10 program using PLAN and EXPLORE as outcomes. Saga 2 is from the 2014-15 school year and is a grades 9-10 program using PLAN and EXPLORE as outcomes.

This measure of tutoring efficiency represents one approach to translating standardized effect sizes (Baird & Pane, 2019; Kraft, 2020), relying on effect sizes of tutoring program impact produced in the context of a well-designed RCT that meets the highest standards of evidence (i.e., ESSA Tier I), and therefore on causal estimates of tutoring program impact. This measure also relies on grade-specific benchmarks, such as those based on the average annual gain in effect sizes (by grade and subject area) from nationally-normed assessments of student achievement (Hill et al., 2008). For the purposes of informing decisions about the relative effectiveness of tutoring programs, tutoring efficiency enables a straightforward comparison of program impact across tutoring providers in units that have intuitive appeal and which account for variation in the average annual growth in student learning in different grades and subject areas.

Despite the utility of our tutoring efficiency measure for facilitating straightforward comparison across program impact estimates, three limitations warrant discussion and future exploration.

First, this measure does not take into account differences in the counterfactual condition; namely, the educational services received by students in the control group in lieu of receiving tutoring, such as business-as-usual instruction, Tier 2 supports, or Tier 3 supports, which likely influences the magnitude of the program impact of a given tutoring program or provider. For example, if the counterfactual condition for an early literacy tutoring program is independent play in one study and a supplemental literacy block in a second study, then differences in program impact would likely be expected given differences in the comparison condition.

Second, our measure of tutoring efficiency does not take into account potential differences in the average annual growth in student learning among students included in the study sample. Specifically, the annual learning growth among students included in an experimental evaluation's study sample (and who tend to most need and benefit from tutoring services) may, on average, differ from the average annual growth in student learning documented in Hill et al. (2008), which relies on nationally representative samples of students (i.e., national norming samples across seven standardized exams in reading/literacy and six standardized exams in math).

Third, a given tutoring program's efficiency based on a single RCT may not be generalizable to different schooling contexts. For example, an estimate of tutoring program impact from a single RCT could be highly efficient, but may not replicate when that same tutoring program is implemented in a different school or district setting. Thus, more RCTs of the same programs in different contexts are necessary to understand the replicability of a given tutoring provider's efficiency. Ongoing replication would also enable the incorporation of multiple impact estimates from multiple schooling settings for a single provider into a weighted average measure of tutoring efficiency. For example, a weighted efficiency measure could, for a single tutoring program, incorporate both variation in program impact estimates and variation in the number of treated students from multiple RCTs.

Section Three

Proposing a Measure of the Cost Effectiveness of Tutoring

A measure of tutoring efficiency provides a policy-relevant estimate of the total number of hours of tutoring necessary to improve student learning by one month. Yet, there is significant heterogeneity across program providers in the cost of tutoring services; estimates of the costs associated with high-dosage tutoring programs range from $1,000 to more than $3,000 per pupil. For example, programs that tutor students in groups tend to have lower per pupil costs than those that provide one-on-one tutoring. Similarly, the training and qualifications of tutors recruited by a program tend to be correlated with tutor wages, which in turn influences the per pupil costs of tutoring. This means that tutoring providers with lower costs can provide more hours of tutoring than higher-cost providers. At the same time, even as the cost of tutoring varies across providers, the amount of tutoring (i.e., dosage) might also vary among providers with similar per pupil costs. This complicates efforts to compare tutoring program costs on a cost per hour basis, let alone on a cost effectiveness basis. Thus, since variability in the cost of tutoring leads to variation in the amount and duration of tutoring program dosage, we want a complementary measure that normalizes tutoring program impact by both tutoring dosage and program costs.

To support policymaker and school leader decisions on the efficient allocation of educational resources, we next present a measure of student learning achieved per dollar of tutoring investment. We propose a measure of the cost effectiveness of tutoring program impact, which we define as the additional months of student learning produced at a fixed (per pupil) cost of $1,000. This requires us to first identify the cost factors that contribute to a tutoring provider's programmatic costs. We note here that there is currently limited information in the field about specific tutoring programs to calculate a cost effectiveness measure that relies on existing cost data. So, following Jacob et al. (2018), we define the following categories of programmatic costs: (i) materials, including curriculum and technology; (ii) tutors, including tutor training, professional development, and compensation; (iii) supervisory and/or administrative support; and (iv) other costs such as facilities or transportation.[13]

Further, we wish to distinguish between two margins of programmatic costs: the **cost to school budgets**, which represents the per pupil cost that schools must pay to tutoring providers; and the **cost to society**, which reflects the total (per pupil) cost of the tutoring services, including any in-kind transfers (e.g., philanthropy), public subsidies (e.g., AmeriCorps volunteers and related services to support their participation), and the opportunity costs associated with, for example, volunteer time. For example, while one tutoring provider in our analytic sample (Reading Partners) has a reported per pupil total (societal) cost of $3,610, this program costs schools $710 per pupil due to in-kind contributions from Reading Partners and from non-pecuniary costs associated with volunteer tutors (Parker et al., 2019). Meanwhile, another tutoring provider in our sample (Saga Education) cites total programmatic costs of $3,500 per pupil, but indicates that the cost to schools could be significantly less through the use of AmeriCorps Fellows as tutors (Guryan et al., 2023). A third provider in our sample (Minnesota Reading Corps) cites a cost to schools of $304 per pupil and a total cost to society of $1,690 per pupil (Bowden et al., 2018). It is important to note that these cost estimates rely on different categories of programmatic costs, different methods for collecting and calculating costs, and differ in the degree of detail about how programmatic costs might vary across program sites. This lack of standardization, even when cost-estimates are conducted rigorously, emphasizes the challenge of making cost comparisons across programs.

Next, the protocols and processes by which cost data might be collected from tutoring providers is a key policy question. For example, in states with state-approved lists of tutoring providers (e.g., Arkansas, Ohio), regulations or policy revisions could be implemented to mandate that tutoring providers on state-approved lists provide data on programmatic costs. However, in the vast majority of states that do not have state-approved lists of tutoring providers, how might state and local agencies work with tutoring providers to collect cost-related data? This is a policy area in the tutoring field that requires significantly more attention from policymakers, school leaders, and researchers so that there is greater transparency around tutoring program costs, particularly as tutoring costs are necessarily a constraint for the total hours of tutoring (i.e., dosage) available to schools and students. Ultimately, the collection of programmatic costs - and, in particular, how costs are documented and described - should be standardized across providers and settings to ensure costs are accurately and transparently reported.

[13] Kraft (2020) proposes a schema for interpreting effect sizes from causal studies of educational interventions with standardized achievement outcomes and defines a cost effectiveness ratio as the ratio of program effect size (in SD units) divided by the program's cost per pupil (in dollars). According to Kraft (2020), a per-pupil cost of $500 (or less) may be considered a low cost program; a per-pupil cost of $500-$4000 may be considered a moderate cost program; and a per-pupil cost of greater than $4,000 may be considered a high cost program.

Once tutoring program costs have been collected, a measure of the cost effectiveness of tutoring is possible. We propose the following measure of cost effectiveness that relies on (a) a valid causal estimate of tutoring program impact from a well-conducted RCT; and (b) the cost of tutoring dosage. Specifically, our measure of cost effectiveness requires the following variables: (i) hours of tutoring for an additional month of learning (this is our measure of Tutoring Efficiency); (ii) dosage of tutoring (in hours); (iii) per pupil cost (either to the school or to society); and (iv) hours of tutoring that can be purchased for $1,000 per pupil. Relying on these variables, we then construct a measure of cost effectiveness, which we define as the number of months of learning for $1,000 per pupil cost of tutoring, as follows:

$$(1)\ \textit{Hours of Tutoring for \$1,000 per pupil} = \frac{\textit{Dosage (hours)}}{\left(\frac{\textit{Cost per pupil}}{\textit{\$1,000}}\right)}$$

$$(2)\ \textit{Cost Effectiveness} = \frac{\textit{Hours of Tutoring for \$1,000 per pupil}}{\textit{Hours of tutoring for an Additional Month of Learning}} = \frac{\textit{Hours of Tutoring for \$1,000 per pupil}}{\textit{Tutoring Efficiency}}$$

Cost Effectiveness: Number of additional months of learning gained by investing $1,000 per pupil

$$\text{Cost Effectiveness} = \frac{\text{Hours of tutoring for \$1,000 per pupil}}{\text{Tutoring Efficiency}}$$

Table 4 provides a stylized example of how our proposed measure of the cost effectiveness of tutoring might be calculated. Program A is highly efficient - it requires just 8 hours of tutoring time to improve student achievement by one additional month. However, Program A is costly in absolute terms (at $3,200 per pupil) and provides less tutoring dosage in terms of hours of total tutoring (32 hours). Thus, the return on investment for Program A - 1.3 months of learning at a cost of $1,000 per pupil - is relatively low. In contrast, Program D, which has the same tutoring efficiency as Program A, is significantly less costly (at $500 per pupil) and provides significantly more tutoring dosage (54 hours) compared to Program A. As a result, Program D is significantly more cost effective, increasing student learning by 13.5 months at a cost of $1,000 per pupil. Program B and C have similar cost effectiveness despite very different measures of tutoring efficiency and cost. This stylized example reveals that the tutoring field should aim to identify (and develop) programs that are highly efficient in terms of increasing student achievement and which provide tutoring at a relatively low per pupil cost. It also shows the importance of taking into account both efficiency and cost to properly assess tutoring programs.

Table 4. Stylized Example of the Cost Effectiveness of Tutoring

Program	**Tutoring Efficiency** (hours of tutoring for an additional month of learning)	**Dosage** (hours)	**Cost Per Pupil** ($)	**Hours of Tutoring for $1,000 per pupil**	**Cost Effectiveness** (months of learning for $1,000 per pupil)
A	8	32	$3,200	10	1.3
B	90	90	$500	180	2.0
C	16	54	$1,600	33.8	2.1
D	8	54	$500	108	13.5

Efforts to select the most cost-effective ways to accelerate student learning are a core function of education decision-makers, policymakers, and researchers. However, it is exceedingly uncommon for decision-makers to rely on measures that contextualize and compare the relative (cost) effectiveness of a program or intervention to inform educational input decisions, such as whether to allocate scarce resources towards reducing class size or investing in supplemental instruction. Understanding costs is also challenging due to the variability in the cost structures of education interventions. The aim of our proposed measure of the cost effectiveness of tutoring is to provide decision-makers, policymakers and researchers a valuable tool with which to compare programs to each other even if there are differences in grade, subject, dosage and cost. This measure could also be leveraged to assess the cost effectiveness of educational interventions beyond tutoring. Notably, while this measure of cost effectiveness is straightforward by design, it requires valid cost data in order to accurately measure an educational intervention's cost effectiveness.

Section Four

Known Unknowns: Accelerate's Research Agenda

Despite the extent of existing research on the impact of tutoring, much remains unknown. There is a need for more research into the specific tutoring programs that improve student learning, for which students, in what educational contexts, at what costs, and the programmatic features most associated with student achievement gains. In particular, based on our review of Nickow et al. (2023) and the existing landscape of evidence on tutoring program impact, researchers and policymakers alike would benefit from more studies that address the following:

Understudied Student Groups

- Focus on programs that serve students older than grade 2, especially literacy programs for students in grades 6-12. It is vital to ensure that tutoring programs are available (and accessible) to support students in the upper grades and that these programs are subjected to the same rigorous expectations (and evaluation) as those that provide tutoring to younger students.
- Focus on results for key student groups, such as special education students with IEPs and multilingual learners. Serving students who may need tutoring services the most means making sure tutoring programs work for those particular students.

Study Design

- Meet What Works Clearinghouse and ESSA Tier 1 evidence standards with a minimum of 350 study participants. More studies with larger sample sizes will provide a more complete picture of the impact of tutoring when done at scale, and greater opportunity to disaggregate impacts by student characteristics to explore potentially differential impacts for different groups of students. Larger study samples will also provide more precise estimates of tutoring impacts and reduce uncertainty around the expected impact of tutoring on student learning.
- Incorporate multiple treatments (e.g., 3-armed RCTs with large samples) that enable a direct test of program design features such as tutor type, session length, program length, tutor-student ratio, dosage, and tutor training intensity. Directly testing program features, while also continuing to assess the average impact of tutoring programs more generally, will help the field understand which program design features are most essential to improving student learning.

Study Outcomes

- Focus on policy relevant outcomes such as reading comprehension and/or end-of-year state exams. Demonstrating the impact of tutoring on proximal outcomes provides insight into how specific features of tutoring curriculum might translate into changes in discrete skills. Yet, the extent to which improvements in discrete skills translates to more general knowledge and more policy-relevant outcomes that impact students' life outcomes is an area requiring significantly more research attention.
- Include longer term outcomes to assess the extent to which the impact of tutoring persists (or fades away). A better understanding of how tutoring impacts the long term learning trajectory of students will clarify both the long term cost-effectiveness of tutoring and whether new improvement strategies are needed (such as booster doses of tutoring).
- Link improvements in foundational skills to performance on state-mandated end-of-year exams (e.g., grade 3 reading assessments). Given the renewed focus on developing early elementary students' foundational literacy skills, it is becoming increasingly important to understand whether (and to what extent) improvements in foundational skills translate into broader measures of student performance.

Data Collection and Replication

- Collect more precise, valid, and itemized data on program costs (both to the school/district and to society at large). Understanding program cost and being able to compare costs across different tutoring programs and with other education interventions is a core component of understanding the cost-effectiveness of tutoring and, ultimately, the ability to scale effective tutoring programs and practices.
- Engage in ongoing replication of impact studies of programs that have already completed RCTs. Understanding the generalizability and reliability of a program, as well as its range of potential outcomes, will improve the use and usefulness of tutoring metrics (e.g., tutoring efficiency and cost effectiveness) and will support more informed and nuanced decision-making. Ongoing replication also prevents wholesale judgment of a program based on a single program impact estimate derived from just one experimental evaluation in a particular research setting.

Table 5. Summary of Research Agenda

Understudied Student Groups	• **Grade 3+** | Focus on programs that serve students older than 2nd grade, especially literacy programs for students in grades 6-12 • **Specific student populations** | Including students with IEPs, multilingual students
Study Design	• **Rigorous evidence** | Meet What Works Clearinghouse and ESSA Tier 1 evidence standards (> 350 students) • **Multiple treatments** | Enable direct test of program design features
Study Outcomes	• **Policy-relevant outcomes** | Including reading comprehension and/or EOY state exams • **Longer-term impact** | Assess the extent to which impact of tutoring persists • **Longitudinal approach** | Link improvements in foundational skills to performance on EOY state exams
Data Collection and Replication	• **Cost data** | Collect more precise and itemized data on program costs (both to the school/district and to society at large) • **Replication trials** | Enables insight into the generalizability and reliability of a program's impact across different schooling settings

Summary and Policy Implications

In this research report, we describe and summarize the most rigorous evidence on the impact of tutoring, which shows that tutoring can significantly improve student learning, outperforming a host of educational interventions. Yet, our re-analysis of the evidentiary base reveals that the scope of evidence on tutoring impacts is less robust than initially expected. Of the 89 experimental evaluations of tutoring programs included in Nickow et al. (2023), just 15 would meet the highest standards of evidence on program impact. Thus, significantly more evidence is necessary to identify the tutoring providers that meaningfully improve student achievement for students in different grade levels, with different educational needs, and in different schooling environments. Additional evidence is also needed to determine which features of tutoring program design are associated with improvements in student achievement. Our research agenda aims to ensure that the students who most need academic support are receiving the most efficient and cost-effective educational interventions.

Relying on a sample of well-designed experimental evaluations of tutoring impacts from the existing literature, we propose and construct a measure of tutoring efficiency which contextualizes the impact of tutoring (in terms of grades, subject, and dosage). This efficiency measure is essential for comparing tutoring programs that vary widely in their program design, and offers valuable guidance to policymakers and education leaders about the relative effectiveness of different tutoring programs and providers. At the same time, there is significant heterogeneity across program providers in the cost of tutoring services. And, since our measure of tutoring efficiency does not account for variation in program costs, we propose an approach for identifying relevant data on programmatic costs and a method for calculating the cost effectiveness of tutoring programs. Together, measures of tutoring efficiency and cost effectiveness offer important new accountability metrics and decision-making tools to inform the evaluation and relative effectiveness of tutoring providers.

While high-dosage tutoring effectively improves student outcomes, there remains a great deal that is still unknown about how best to use this intervention. This research report lays out an actionable research agenda and describes new, policy-relevant metrics for use in comparing tutoring providers and for evaluating provider effectiveness. Accelerate strongly supports the need for an expansion of the evaluation work that is currently being done in the tutoring field. We see the need for even more experimental evaluations on many more tutoring providers. This expansion of the evidentiary base would enable insights into variation in tutoring efficiency and cost effectiveness both across and within tutoring providers. Additional experimental evaluations on the same tutoring providers would enable a better understanding of the reliability, generalizability, and range of potential outcomes for a particular provider. And, while this research report offers novel metrics to compare the relative impact of tutoring providers, more work is necessary to more completely understand which tutoring providers work, under what conditions, and for which students. Even though the current landscape of evidence on the impact of tutoring remains insufficient, this research report provides a roadmap for moving the field forward to support informed decision-making.

At Accelerate, we focus on program effectiveness and return on investment (ROI) as the primary factors that inform how education leaders select interventions. School leaders weigh trade-offs not just between tutoring options but also between status quo options. Replacing the status quo with a promising intervention like tutoring requires time, effort, and fiscal resources. Clearer expectations about the likely return on educational investments will empower leaders to make optimal decisions for their schools. The tutoring efficiency and cost effectiveness measures described in this research report offer important new insights into program effectiveness and ROI, and provide new guidance to education leaders working to identify tutoring models that meaningfully improve student achievement and can be scaled to reach significantly more than the 10 percent of students nationally who currently receive high-dosage tutoring.

Policy Recommendations for Key Stakeholders

- **Researchers** should focus on large-sample evaluations to produce rigorous evidence on tutoring effectiveness so that education leaders can make the most cost-effective decisions when selecting tutoring providers.
- **Tutoring providers** should gather data on all students they serve, not just those who receive the full dosage, and should provide more transparency around the cost of their programs.
- **Federal agencies** should lead high-quality research efforts and enact policies that incentivize tutoring providers to meet rigorous standards.
- **States** should identify and curate a list of vetted, evidence-based high-dosage tutoring programs that meet a standard of efficiency and mandate cost reporting for vetted tutoring providers that make it onto state lists.
- **School districts** should build tutoring into their budgets for the long term and rely on evidence to select only the most effective and efficient programs.

References

Baird, M. D., & Pane, J. F. (2019). Translating Standardized Effects of Education Programs Into More Interpretable Metrics. Educational Researcher, 48(4), 217-228.

^Barnes, M. A., Klein, A., Swank, P., Starkey, P., McCandliss, B., Flynn, K., Zucker, T., Huang C., Fall, A. & Roberts, G. (2016). Effects of tutorial interventions in mathematics and attention for low-performing preschool children. Journal of Research on Educational Effectiveness, 9(4), 577-606.

Bowden B., Escueta M., Muroga A., Rodriguez V., & Levin H.M. (2018). Report for Minnesota Reading Corps Pre-K Program Cost Analysis. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Carbonari M. V., Davison M., DeArmond M., Dewey D., Dizon-Ross E., Goldhaber D., Hashim A., Kane T. J., McEachin A., Morton E., Patterson T., Staiger D.O. (2022). The Challenges of Implementing Academic COVID Recovery Interventions: Evidence from the Road to Recovery Project. CALDER Working Paper No. 275-1222.

*Clarke, B., Doabler, C. T., Kosty, D., Kurtz Nelson, E., Smolkowski, K., Fien, H., & Turtura, J. (2017). Testing the efficacy of a kindergarten mathematics intervention by small group size. AERA open, 3(2).

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

#Fryer R.G. Jr. & Howard-Noveck M., 2020. High-Dosage Tutoring and Reading Achievement: Evidence from New York City. Journal of Labor Economics, University of Chicago Press, 38(2), 421-452.

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

*Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., Farkas, G., Fryer, R. G., Jr., Mayer, S., Pollack, H., Steinberg, L., & Stoddard, G. (2023). Not too late: Improving academic outcomes among adolescents. American Economic Review, 113(3), 738–765.

Hill, C., Bloom, H., Black, A. & Lipsey, M. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. Child Development Perspectives. 2. 172 - 177.

*Jacob, R., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. Journal of Research on Educational Effectiveness, 9(sup1), 67-92.

Jacob R., Erickson A., & Mattera S. (2018). Launching Kindergarten Math Clubs: The Implementation of High 5s in New York City. MDRC.

*Jacob R., Erickson A. & Mattera S. (2020). Evaluating the Impact of Small Group Supplemental Math Enrichment in Kindergarten, Journal of Research on Educational Effectiveness, 13:3, 381-407.

References

#Jones C. J. (2018). Future forward early literacy: Testing the impact of a family-school-community partnership literacy intervention. School Community Journal, 28, 247–264.

#Jones C.J. & Christian M. (2021). The Results of a Randomized Control Trial Evaluation of the SPARK Literacy Program: An Innovative Approach that Pairs One-on-One Tutoring with Family Engagement, Journal of Education for Students Placed at Risk (JESPAR), 26:3, 185-209.

+Jones, C. J., Reeves, M., Li, D., & Gilman, L. (2023). What Is the Sustained Impact of Future Forward on Reading Achievement, Attendance, and Special Education Placement 5 Years After Participation? Educational Evaluation and Policy Analysis.

Kuhfeld M., & Lewis K. (2022). Student achievement in 2021–2022: Cause for hope and continued urgency. NWEA Research.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. Educational Researcher, 49(4), 241-253.

Kraft, M. A., & Falken, G. T. (2021). A Blueprint for Scaling Tutoring and Mentoring Across Public Schools. AERA Open, 7.

*Lee, Y. S., Morrow-Howell, N., Jonson-Reid, M., & McCrary, S. (2012). The effect of the Experience Corps® program on student reading outcomes. Education and Urban Society, 44(1), 97-118.

+Loeb S., Novicoff S., Pollard C., Robinson C., & White S. (2024). The Effects of Virtual Tutoring on Young Readers: Results from a Randomized Controlled Trial. National Student Support Accelerator.

+Markovitz, C. E., Hernandez, M. W., Hedberg, E. C., & Silberglitt, B. (2014). Impact Evaluation of the Minnesota Reading Corps K-3 Program. Corporation for National and Community Service.

#Markovitz, C. E., Hernandez, M. W., Hedberg, E. C., & Whitmore, H. W. (2022). Evaluating the Effectiveness of a Volunteer One-on-One Tutoring Model for Early Elementary Reading Intervention: A Randomized Controlled Trial Replication Study. American Educational Research Journal, 59(4), 788-819.

^Miller, S., Connolly, P., & Maguire, L. K. (2012). The effects of a volunteer mentoring programme on reading outcomes among eight-to nine-year-old children: A follow up randomized controlled trial. Journal of Early Childhood Research, 10(2), 134-144.

^Miller, S., & Connolly, P. (2013). A randomized controlled trial evaluation of time to read, a volunteer tutoring program for 8-to 9-year-olds. Educational Evaluation and Policy Analysis, 35(1), 23-37.

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. American Educational Research Journal, 61(1), 74-107.

*Parker, D. C., Nelson, P. M., Zaslofsky, A. F., Kanive, R., Foegen, A., Kaiser, P., & Heisted, D. (2019). Evaluation of a math intervention program implemented with community support. Journal of Research on Educational Effectiveness, 12(3), 391-412.

References

+Patel R., Jabin N., Bussard L., Cartagena J., Haywood S. & Lumpkin M. (2017). Switch-on Effectiveness Trial. Education Endowment Foundation.

^Rebok, G. W., Carlson, M. C., Glass, T. A., McGill, S., Hill, J., Wasik, B. A., ... & Rasmussen, M. D. (2004). Short-term impact of Experience Corps® participation on children and schools: Results from a pilot randomized trial. Journal of Urban Health, 81(1), 79-93.

Rich, M. (May 29, 2013). In Raising Scores, 1 2 3 is Easier Than A B C. The New York Times. Retrieved from: https://www.nytimes.com/2013/05/30/education/reading-gains-lag-improvements-in-math.html

^Ritter, G., & Maynard, R. (2008). Using the right design to get the 'wrong' answer? Results of a random assignment evaluation of a volunteer tutoring programme. Journal of Children's Services.

Robinson, C., Kraft, M., Loeb, S. & Schueler, B. (2021) Design Principles for Accelerating Student Learning With High-Impact Tutoring. EdResearch for Action.

*Sirinides, P., Gray, A., & May, H. (2018). The Impacts of Reading Recovery at scale: Results from the 4-year i3 external evaluation. Educational Evaluation and Policy Analysis, 40(3), 316-335.

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

+Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. Journal of Educational Psychology, 91(4), 579–593.

*Studies included in the re-analysis analytic sample that were also included in Nickow et al. (2023).

#Studies included in the re-analysis analytic sample that were not included in Nickow et al. (2023).

^Studies that were included in Nickow et al. (2023) with a study sample greater than 350 but not included in the re-analysis analytic sample because they did not meet all selection criteria.

+Studies that the authors reviewed that were not included in Nickow et al. (2023) and which were not included in the re-analysis analytic sample because they did not meet all selection criteria.

Accelerate
The National Collaborative for Accelerated Learning