Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1031

What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability

Matthew A. Kraft
Brown University

Beth E. Schueler
University of Virginia

Grace Falken
Brown University

U.S. public schools are engaged in an unprecedented effort to expand tutoring in the wake of the COVID-19 pandemic. Broad-based support for scaling tutoring emerged, in part, because of the large effects on student achievement found in prior meta-analyses. We conduct an expanded meta-analysis of 265 randomized controlled trials and explore how estimates change when we better align our sample with a policy-relevant target of inference: large-scale tutoring programs in the U.S. aiming to improve standardized test performance. Pooled effect sizes from studies with stronger target-equivalence remain meaningful but are only a third to a half as large as those from our full sample. This result is driven by stark declines in pooled effect sizes as program scale increases. We explore four hypotheses for this pattern and document how a bundled package of recommended design features serves to partially inoculate programs from these attenuated effects at scale.

VERSION: October 2024

Suggested citation: Kraft, Matthew A., Beth E. Schueler, and Grace Falken. (2024). What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability. (EdWorkingPaper: 24-1031). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/zygj-m525

What Impacts Should We Expect from Tutoring at Scale?
Exploring Meta-Analytic Generalizability

Matthew A. Kraft
Brown University

Beth E. Schueler
University of Virginia

Grace T. Falken
Brown University

October 2024

Abstract

U.S. public schools are engaged in an unprecedented effort to expand tutoring in the wake of the COVID-19 pandemic. Broad-based support for scaling tutoring emerged, in part, because of the large effects on student achievement found in prior meta-analyses. We conduct an expanded meta-analysis of 265 randomized controlled trials and explore how estimates change when we better align our sample with a policy-relevant target of inference: large-scale tutoring programs in the U.S. aiming to improve standardized test performance. Pooled effect sizes from studies with stronger target-equivalence remain meaningful but are only a third to a half as large as those from our full sample. This result is driven by stark declines in pooled effect sizes as program scale increases. We explore four hypotheses for this pattern and document how a bundled package of recommended design features serves to partially inoculate programs from these attenuated effects at scale.

Author Note: Correspondence regarding the manuscript can be sent to Matthew Kraft at mkraft@brown.edu.
This research was generously supported by the National Science Foundation and the William T. Grant Foundation. We are grateful for research assistance from Luyan Nguyen, Samuel Lynch, Hannah Sexton, Virginia Lovison, Sara White, Marlene Almanzar Garcia, Audrey Whitten, Leo Gordon, Halle Bryant, Jay Philbrick, Deena Haque, Tommy Bellaire, Summer Dain, Stephanie Tu, Minerva Lopez, Mary Lau, Charlotte DeVaughn, Eli Talbert, Margaret Brehm, Meagan Thompson, Isabelle Saillard, Gabrielle Oliver, among others at Brown University and the University of Virginia. We also appreciate feedback from Elizabeth Tipton, Matthew Steinberg, Alan Safran, along with participants in the Program on Education Policy and Governance Colloquium at the Harvard Kennedy School, the National School Support Accelerator Annual Conference at Stanford University, the Society for Research on Educational Effectiveness Annual Conference, and the Syracuse University, Maxwell School lecture series.

Introduction

We are living in a rare moment where a collective effort is underway to change one of the core organizing principles of modern schooling. Historically, education – both formal and informal – was primarily an individualized endeavor with tutors and pupils or master craftsmen and apprentices working together, one-on-one. The rise of large-scale public education systems over the last two centuries evolved around a different organizing principle – one in which teachers became charged with the task of educating entire classrooms of students (Tyack, 1974). While teaching students in groups allowed these systems to expand access rapidly, it also created substantial challenges for educators to meet the full spectrum of students' individual needs.

The COVID-19 pandemic toppled the precarious balance teachers have long tried to achieve between whole-class instruction and differentiated instruction. The public health crisis caused widespread school closures as well as acute hardships for many families. In the U.S., researchers estimate that median student achievement fell 0.24 standard deviations (SD) in math and 0.13 SD in reading, with even larger declines among low-achieving students (Callen et al., 2024). The pandemic both exacerbated longstanding inequalities in student achievement and created a shared priority to accelerate learning. This crisis caused the historical pendulum to start swinging back towards individualized instruction as a means of meeting the needs of all students.

In the months following the pandemic, a rare consensus emerged among policymakers, researchers, and practitioners that tutoring had a critical role to play in addressing the educational harms caused by COVID-19. Integrating tutoring into the public education system at scale has become a primary policy response to pandemic-related learning disruptions. Unlike previous unfunded attempts to scale tutoring, such as President Clinton's America Reads initiative, the federal government and individual states catalyzed these efforts with substantial financial investments (NSSA, 2023). The federal Elementary and Secondary School Emergency Relief Fund (ESSER) provided $190 billion to public schools and required districts to spend a sizable fraction of this on student learning acceleration (including 20% of the third wave of ESSER funding) (Goldhaber & Falken, 2024). One estimate projects that states and districts have so far spent over $3 billion of this aid on tutoring (DiMarco & Jordan, 2022).

Efforts to scale tutoring after COVID-19's onset appear to have substantially expanded access to individualized instruction in U.S. public schools. The nationally representative School Pulse Survey found that by December of 2022, 37% of schools reported offering high-dosage tutoring, defined as "Tutoring that takes place for at least 30 minutes per session, one on one or in small group instruction, offered three or more times per week, is provided by educators or well-trained tutors, [and] aligns with an evidence-based core curriculum or program." This statistic increases to 59% when schools were asked if they offer more standard tutoring defined as a less intensive and structured approach to individualized instruction. At the same time, districts have yet to implement these programs at the scale or dosage many believe is required to support a full academic recovery (Goldhaber et al., 2022). Only 20% of schools that reported offering high-dosage tutoring (and 15% that offered standard tutoring) strongly agreed that they were able to effectively provide tutoring to all students in need.

Efforts to expand access to tutoring were, in many ways, evidence-based policy. Meta-analyses conducted by several independent research teams that reviewed randomized controlled trials (RCTs) of tutoring programs have all found large effects of tutoring on test-based measures of achievement in the range of 0.3 to 0.4 SD (Dietrichson et al., 2017; Fryer, 2017; Inns et al., 2019; Nickow et al., 2020, 2024; Pellegrini et al., 2021). These effects are roughly equal to an 11 to 15 percentile point increase, or the amount of learning in reading that upper elementary students in the U.S. typically make in an entire school year (Hill et al., 2008). These impressive findings played a central role in motivating calls by policymakers and researchers – including ourselves (Kraft & Falken, 2021; Robinson et al., 2021) – to advocate for scaling tutoring. Influential technologists such as Mark Zuckerberg and Sal Khan have extolled the moonshot-like potential of tutoring, evoking the eye-popping 2 SD effects found in small-scale studies conducted by University of Chicago doctoral students under the supervision of Benjamin Bloom in the 1980s. However, scholars have recently raised new critiques about Bloom's 2-sigma studies (Barnum, 2018; von Hippel, 2024) and the generalizability of pooled effect sizes generated from meta-analytic reviews (Dahabreh et al., 2020; Littell, 2024; Slough & Tyson, 2023).

In this paper, we conduct an expanded and updated meta-analysis of RCTs evaluating tutoring programs to explore the external validity of pooled effect size estimates. The common empirical focus on RCTs bolsters the internal validity of meta-analytic estimates, allowing researchers to draw credible inferences about the causal impacts of tutoring programs on student outcomes. However, despite combining findings from a wide variety of settings, samples, and treatments, meta-analytic reviews of experimental studies with small to medium non-probability samples do not necessarily produce estimates that generalize to broader efforts to scale tutoring (Littell, 2024). As many scholars have highlighted, strong internal validity does not beget broad external validity (Banerjee & Duflo, 2009; Esterling et al., 2024; Pritchett & Sandefur, 2015).

We seek to answer the question: What expectations should we have for tutoring effects on standardized test scores for large-scale programs implemented in the U.S.? We address this question by generating pooled effect sizes from a sample of 265 RCTs published between 1967 and 2023 and examining the sensitivity of our results to sample restrictions that better align our estimates with a specific, policy-relevant target of inference: large-scale tutoring programs in the U.S. aiming to improve achievement on standardized tests. Consistent with prior meta-analyses, we find a large, pooled effect size of 0.42 SD on student achievement across our full sample. These effects are driven, in part, by the strikingly large effects of literacy tutoring programs in elementary grades (0.46 - 0.48 SD), which constitute 73% of the estimates in our sample.

Our analyses reveal a stark pattern of declining effects of tutoring programs when taken to scale. When we restrict our sample to larger-scale tutoring programs implemented in the U.S. and evaluated based on third-party standardized assessments, our pooled estimates shrink to a third to a half the size of our unrestricted estimates. This difference is largely due to restricting the sample to large-scale programs. In our preferred analytic samples, we estimate a pooled effect size of 0.21 SD for programs serving 400 to 999 students and 0.16 SD for programs serving 1,000 students or more. We view these effects both as more plausible for large-scale programs and as still having considerable policy importance given their meaningful magnitude and strong external validity. Still, such effects for tutoring at scale are far from guaranteed. We observe considerable variation across individual tutoring program effects, and estimates from quasi-experimental studies of programs serving thousands of students are often even smaller.

We then explore several hypotheses that might explain the pattern of declining effects with scale – a widely documented phenomenon in the broader education research literature (Cheung & Slavin, 2016; Kraft, 2020, 2023). We find mixed evidence that the declining results are an artifact of selective reporting due to publication bias or *p*-hacking. It is possible, however, that the common practice to use the within-sample standard deviation to estimate effect sizes causes these estimates to be artificially larger for studies of smaller programs serving more targeted and homogenous populations. We also find evidence that some tutoring program design features differ systematically across program size, with increasing student-teacher ratios and declining dosage. We suspect another possible explanation for declining effects with scale is that tutor effects are heterogeneous across students, causing the marginal effect of tutoring to decline as programs expand to serve more students that stand to benefit somewhat less. Finally, we also find evidence from recent studies that implementation quality often declines at scale.

Encouragingly, we do find that a combination of tutoring program design features identified in the research literature as best practices somewhat buffers against the large decline in effects we find at scale. However, schools and districts are often motivated to expand tutoring while operating within budget constraints. This creates a tension between maintaining fidelity to best practices and supporting more students. We conclude by examining how common approaches to addressing scaling challenges, such as high program costs and limited tutor supply, might affect program efficacy at scale. We find generally inconclusive evidence about the ability to maintain program efficacy when tutoring is delivered online, when student-tutor ratios are increased, and when dosage is decreased. We find suggestive evidence that peer tutoring may offer a cost-effective model for scaling.

Our study makes several contributions to literature. We extend prior tutoring meta-analyses by compiling a sample of 265 RCTs, roughly three times the number of studies as the largest prior reviews. This large sample allows us to explore how our overall effect size estimates compare to those for sub-samples of studies that are more aligned to the target of inference used by many researchers and policymakers. Second, our study also serves as an applied example of why it is critical to attend to external validity when conducting meta-analytic reviews and engaging in evidence-based policymaking. Finally, our analyses generate important insights to inform ongoing efforts to take tutoring to scale within the U.S. public school systems in a sustainable way. Our findings provide stronger, more externally valid evidence to support investments in tutoring, while also recalibrating expectations towards more plausible gains for students.

Methods

Literature Search Procedures

We began by searching for articles in seven electronic databases, including Academic Search Premier, APA PsychInfo, AEA EconLit, ERIC, Google Scholar, Science Direct, and Web of Science. We also searched two working paper series, from the Brown University Annenberg Institute and the National Bureau of Economic Research, to ensure we captured studies not yet published in peer-reviewed outlets (Alexander, 2020; Pigott & Polanin, 2020). Searching this range of sources was essential to minimize the extent to which we were missing key research, especially work produced by scholars from historically marginalized groups (Boveda et al., 2023). Our search terms included keywords related to (a) tutoring (e.g., “tutor”), (b) educational contexts (e.g., “school”), and (c) impact evaluation research methods (e.g., “RCT”). We used Boolean operators between all terms, specifically “OR” between terms within each of these three keyword categories, many of which were synonyms, and “AND” between each of the three categories to maximize the relevance of search results without overlooking key studies. We identified 45 preexisting reviews and meta-analyses of tutoring-related interventions and scanned the reference lists of these for new studies. We supplemented this search by monitoring social media and email newsletters from research centers. We continued our literature search through the end of 2023, the cutoff date for studies we formally coded. Though we stopped coding new studies, we continued to track newly released studies and incorporate several into our narrative synthesis and discussion. Our search generated over 14,000 studies. After removing duplicates, we followed Pigott and Polanin (2020) and had two team members conduct an initial screening for relevance using titles and abstracts. This left 1,347 studies that we subjected to an in-depth inclusion review of the full texts, ultimately resulting in a final analytic sample of 265 studies.

Inclusion Criteria

To identify our analytic sample, we assessed studies against eight inclusion criteria: 1) human tutoring, 2) 1:1 or in small groups, 3) focused on academics, 4) measured effects on standardized tests in math or reading, 5) K-12 students, 6) in an OECD country, 7) RCT design, and 8) randomized more than 20 students or 4 classrooms. First, programs under study needed to meet a broad definition of tutoring: one non-parental individual providing academic support to a single student or small group of students. We excluded studies of individualized instruction provided by a book, computer program, or other curricular tool without the direct support of a human tutor. While we included studies of programs where the tutor was a teacher, paraprofessional, college student, volunteer, or peer, we excluded studies of parent tutoring programs because all relevant studies we identified evaluated models of parent training or professional development rather than direct parent-child instruction. Second, the tutoring intervention must have been implemented with either a 1:1 student-tutor ratio or in groups of 8 or fewer students.[1] Third, the tutoring content had to focus on academic subjects. This excluded, for example, studies of mentoring or socioemotional interventions without an academic component. Fourth, our focus on academic interventions also meant that studies needed to report effects on academic outcomes, specifically standardized tests and researcher-generated assessments measuring performance in either reading or math. We excluded studies where the only outcome was a non-test academic measure (e.g., GPA, attendance) because the sample of these studies

[1] We recognize not everyone would characterize instruction in groups of 5-8 as an authentic tutoring program. We include such programs because authors commonly applied this term and because it allowed us to cast a wide net.

was too small to facilitate broad comparisons. Fifth, the tutees had to K-12 students. This excluded studies of tutoring in early childhood settings, of college or graduate students, and of adults. Sixth, the intervention had to take place in a member country of the Organization for Economic Co-operation and Development (OECD) given our primary focus was on informing policy in the U.S. (a high-income nation). Seventh, we limited our sample to RCT designs to parallel prior reviews and given RCTs' relative advantage at isolating causal impacts. That said, we supplement our meta-analysis with a synthetic review of recent quasi-experimental studies, which helps us consider tutoring impacts at a scale not captured by most RCTs. Finally, the studies had to have a sample size of more than 20 students when randomization occurred at the student level, or more than four classrooms or schools when randomization was at the classroom or school level.

We also applied inclusion criteria to the effects reported and coded all qualifying estimates from each study. First, the effect estimates had to examine the same outcome as the subject of the tutoring (i.e., we dropped estimates of the impact of math tutoring on reading achievement). Second, we focused on treatment-control contrasts that isolated tutoring whenever possible, dropping estimates where the control condition involved tutoring-like programs and comparisons between treatment arms without a pure no-tutoring control group. However, we included studies in which we judged tutoring to be a key element of a larger set of interventions and reforms that together were evaluated against a business-as-usual control group. Finally, we prioritized estimates from reduced form models that capture the effect of offering tutoring. We view these intent-to-treat estimates as the relevant impact for the types of inferences policymakers often make about what the effect of a program will be as implemented at scale.

Coding Procedures

Our research team of 20 coders double coded each study in our sample. Coders were trained on a common set of studies until they achieved a consistently high agreement rate with master codes created by our most experienced coders. After coding each study independently, coders then met to reconcile any differences and arrive at a final set of codes. When a pair of coders felt that the reconciliation was not straightforward, they brought questions to the principal investigators for a final determination. The team kept a record of decision rules that resulted from these meetings to ensure consistency across coders and over time.

Our codebook included 128 codes that we grouped into five categories. Some codes varied at the study-level while others varied at the intervention- or estimate-level. The first group of codes catalogued study information such as publication type (e.g., peer-reviewed article, working paper) and publication year. The second group tracked information about the context in which the study occurred such as the country, school level, and participant demographics. The third set covered information about the intervention itself and the treatment/control contrast (e.g., student-tutor ratio, the dosage, tutor type). The fourth category was information on the methods used by the study's authors such as the level of assignment to treatment, whether standard errors were clustered at the appropriate level, and whether we had concerns about attrition or contamination of the randomization. The fifth set included information about the effects, including estimated effect sizes, standard errors, sample sizes, and outcome instruments.

We highlight one key code that we use throughout our analyses: the number of treated students. Prior meta-analytic reviews often explore how effect sizes vary by the total sample size of an evaluation. We take a somewhat different approach given our focus on identifying studies that are more closely aligned to a specific target of inference. We code the number of students randomly assigned to receive treatment as an estimate of the number of treated students.

Although we are conceptually interested in the actual number of students who participated in tutoring, this quantity was not consistently reported across studies. Thus, our code for the number of treated students serves as an upper bound of the actual size of the tutoring program.

Calculating Effect Sizes

Study authors reported treatment effects in a variety of ways. Whenever they were available, we defaulted to relying on standardized effect sizes generated from linear regressions estimating standardized mean differences between the treatment and control group, often controlling for baseline covariates. One advantage of model-based estimates is that the associated standard errors typically account for the ways data may be clustered, as recommended by Hedges (2007). When these estimates (and/or their associated standard errors) were unstandardized, we standardized them using unadjusted pre-treatment control group SD whenever possible (if unavailable, we used pooled SD). In other cases, we estimated a standardized effect size using the pre-post treatment means, SD, and sample sizes for the treatment and control group. For each estimate, we then calculated a Hedges' g effect size, correcting for upward bias present for small-sample studies (Borenstein et al., 2009) as follows:

$$g^* = \left(1 - \frac{3}{4(n_T + n_C) - 9}\right) g$$

Here, g^* is the corrected effect size estimate, n_T is the number of treated units (i.e., students or classrooms), n_C is the number of comparison units, and g is the uncorrected effect size estimate.

Meta-Analytic Approach

We generated our pooled standardized effect size estimates using robust variance estimation (RVE) meta-analytic methods (Hedges et al., 2010). Like other meta-analytic techniques, this approach up-weights effects estimated with greater precision, but RVE is unique in that it also accounts for the nesting of impact estimates within clusters. We often observe multiple estimates for a given study (for example, when there are multiple outcomes or interventions examined in a single study) and therefore model this most prevalent type of dependency, as recommended by Tanner-Smith and Tipton (2014). We fit the following model:

$$Y_{ij}^{k} = \beta_0^k + u_j^k + \varepsilon_{ij}^k$$

where Y_{ij}^{k} represents an impact estimate *i* on outcome *k* (either math, reading, or stacking subjects together in a single analysis) from study j. β_0^k is the overall weighted average impact of tutoring on outcome *k* and u_j^k is a study-level random effect. ε_{ij}^k is the residual of a specific effect size estimate from the average effect within its study. In addition to pooled effect estimates and associated standard errors, we also report prediction intervals for select estimates to describe the degree of heterogeneity in our sample and to illustrate the range of plausible effects policymakers might expect for an individual tutoring program (Borenstein et al., 2017).

Towards More Credible Estimates of Program Design Feature Effects

Researchers have typically explored the relative importance of various program features by comparing the pooled effect sizes of tutoring programs with different features. This approach is limited, however, because program features are often bundled and could be correlated with unobserved aspects of program quality (Tipton et al., 2023). We attempt to reduce these potential biases using meta-regressions to examine which moderators predict larger impact estimates, conditional on other study and program design features. We estimate the following model:

$$Y_{ij}^{k} = \beta_0^k + \Gamma X_{i/j} + u_j^k + \varepsilon_{ij}^k$$

Here, we include a vector of study and intervention features ($\Gamma X_{i/j}$). While this model does not allow us to isolate the causal impact of a particular intervention feature on student outcomes, it does allow us to tease apart which of the observable study and intervention characteristics are driving the largest differences in effect size estimates. When possible, we complement these analyses with results from multi-arm RCTs that randomly assign students to tutoring programs that differ only by a single design feature. These studies provide credible causal estimates of the effect of specific program design features but are often underpowered to detect small to medium differences in effects produced by modifying only one aspect of a tutoring program.

Target of Inference

Our aim is to draw inferences about tutoring programs that are most relevant for the target, context, outcomes, and scale of tutoring programs envisioned by U.S. policymakers. Specifically, we hope to inform the expectations of leaders who are seeking to address overall declines and growing gaps in academic outcomes post-COVID by integrating tutoring into the U.S. K-12 public school system. We imagine that because leaders are being held accountable for results on statewide standardized exams that assess a broad set of basic skills, policymakers will be more interested in tutoring impacts on standardized exams that measure general skills as opposed to assessments that measure narrower sets of skills or that are designed by researchers to align tightly with the focal content of the tutoring intervention. Furthermore, our U.S.-centric policy focus makes studies conducted in the U.S. likely to be the most relevant for our target of inference.

A central motivation for our work is to inform efforts to significantly expand access to tutoring programs. We therefore aim to draw inferences about reasonable expectations for the impacts of tutoring programs implemented at scale, as opposed to small-scale pilot programs. Throughout the paper we present estimates of pooled effects sizes across four bins of program size: 0-99, 100-399, 400-999, and 1,000 or more students. We can approximate the relevance of these bin sizes for tutoring programs through a back-of-the-envelope calculation.

Survey data from the "School Pulse Panel" (SPP) show that among districts offering high-dosage tutoring, approximately 28% of district students participate (NCES, 2024). This likely represents a lower bound estimate of the total need for high-dosage tutoring, given that 37% of 4th graders scored below basic on the 2022 National Assessment of Education Progress exam in reading and 38% of 8th graders scored below basic in math. Assuming districts target an average of 28% of students for tutoring, these bins partition public school districts at the 22nd, 57th, and 80th percentiles of district size.[2] Importantly, even though these bins capture a similar number of districts, the number of students in each bin is highly skewed towards the larger program sizes. Districts in the 0-99 program size bin serve 1% of all public school students, 100-399 8%, 400-999 15%, and 1,000+ 76%. Although only 20% of districts might intend to build tutoring programs that serve 1,000 students or more, roughly three-fourths of all public school students attend districts of this size, making it a policy-relevant focus of our analysis.

Findings

Characteristics of Included Studies

Our final analytic sample includes 265 RCTs that evaluate 340 distinct tutoring interventions. We present characteristics of these studies at the study/RCT-level in Table 1. Our sample skews towards recent research with almost two-thirds of included reports published in the years since 2009 and almost 80% in the last 20 years. Only five studies in our sample assess interventions implemented since the beginning of the pandemic, almost all of which provided remote tutoring, giving us limited power to disentangle virtual delivery from the pandemic context. Three-fourths of studies in our sample are peer-reviewed journal articles. The modal study examined a tutoring program in an urban, public school setting.

[2] Corresponding district size for the four bins is 1-357; 358-1,428; 1,429-3,571, and 3,572 or more students.

Our sample reflects substantial imbalance in the subject, grade-level, and size of tutoring programs evaluated in the literature, as illustrated by the evidence gap maps shown in Figures 1 and 2 (Polanin et al., 2023). Most of the studies assess literacy tutoring among early elementary school students (51%) and programs serving fewer than 100 students (59%). This concentration on small elementary reading programs is worth noting because if impacts differ across grade-levels, subjects, or with program scale, pooled results based on our full sample may not be immediately generalizable to other program types.

We provide further details on the characteristics of the programs evaluated in each of these studies in Table 2. Most interventions were delivered in-person (97%), at school (86%), during school hours (76%), using a 2:1 student-tutor ratio or less (62%), and with a provided curriculum (89%). Although individual tutoring was the modal approach (46%), student-tutor ratios varied widely across the sample. We observe greater variation in design choices across the features of tutor type, dosage, and whether students were pulled out of class for tutoring.

Full Sample Estimates of Tutoring Impacts

Similar to prior tutoring meta-analyses, we find notably large, pooled effect sizes across our full sample of studies. As shown in Table 3, we estimate that the average effect on student achievement of a broad variety of tutoring interventions subjected to rigorous evaluation via RCTs is 0.42 SD when stacking math and reading achievement impacts. The associated prediction interval ranges from -0.31 SD to 1.16 SD, illustrating the considerable heterogeneity of impacts we might expect across individual tutoring programs. This large average effect is driven, in part, by the pooled effects of literacy tutoring in lower and upper elementary grades of 0.46 and 0.48 SD, respectively, which make up a large portion of our sample (84%). That said, the pooled effects of tutoring on math achievement are still quite large (0.39 SD). We find inconsistent patterns in tutoring effect size across schooling levels by subject. Impacts of reading tutoring for elementary school students are substantially larger than for middle and high school students. In math, we find the largest effects at the high school level (0.55 SD) followed by upper elementary (0.44 SD). However, for both subjects, we only observe a small sample of high school program effects (13 estimates for math and 27 for reading). The magnitudes of effects are still moderate to large for the school levels and subjects with the smallest pooled effects (0.33 SD for lower elementary math and 0.16 SD for high school reading).

Sensitivity Analyses

We next explore whether our pooled estimates are robust to a variety of sensitivity checks in Table 4. First, we examine whether results differ for studies that may have lower internal validity due to quality concerns with the randomization design or empirical analyses. For example, some authors described their methods as an RCT but indicated or intimated that students, teachers, parents, or administrators had some influence over whether a student ended up in the treatment or control group. Another example is when a sizable number of students were excluded from the analytic sample because of non-compliance, attrition, a move, or some other reason. When we separately examine results based on studies for which we did not have quality concerns, results remain essentially unchanged. Our results are slightly sensitive to excluding outlier effects. When we omit the top and bottom 2.5% of effect size observations the pooled effect size estimate drops to 0.38 SD, a decline that is largely driven by a reduction in pooled reading effects from 0.44 to 0.37 SD.[3]

Finally, we examine whether estimates vary by the decade in which they were published as a rough proxy for study quality. Education research has taken major leaps in terms of

[3] The range of the lowest 2.5% of estimates was -1.70 to -0.34 SD. For the highest 2.5% of estimates, we observe a range of 2.24 to 8.06 SD.

methodological rigor and quality standards over the past three decades, particularly in applying causal inference methods (Angrist, 2004).[4] As shown in Table 4, we find substantial variation in the magnitude of the pooled impacts based on publication decade, with larger estimates prior to 2000 (0.45 SD) and between 2000 and 2009 (0.58 SD) than for those published between 2010 to 2019 (0.38 SD). We observe the smallest impacts for the most recent studies published in 2020 or after (0.27 SD), though even this is large in magnitude. We cannot definitively disentangle whether this variation in impacts is due to methodological changes, policy changes, or other study or program characteristic changes over time, but differences across decades remain even after we control for a host of study and program characteristics, as shown in Table 8.

What expectations should we have for tutoring effects at scale?

Evidence from our meta-analysis of experimental studies. In Table 5, we explore how our pooled effect size estimates change when we restrict our sample to more closely approximate our target of inference. Removing estimates that rely on assessments designed by the research team induces a modest 0.07 SD decline in our aggregate estimate. Removing studies conducted outside the U.S. only trivially reduces our pooled estimate by 0.03 SD. However, restricting the sample to studies that provided tutoring to incrementally larger groups of students profoundly changes the magnitude of our estimates. Using our full sample, we find that programs offering tutoring to fewer than 100 students have a pooled effect size of 0.55 SD, whereas programs tutoring between 100 and 399 students have a pooled effect size of 0.32 SD. As shown in Figure 3, this estimate continues to decline – almost linearly – as we further restrict the sample such that

[4] Another reason we chose to examine effects by decade is because the decades roughly correspond to major periods of education policy development, with the pre-2000 studies representing the period prior to universal test-based accountability, the 2000 to 2009 period representing the No Child Left Behind (NCLB) era, the time from 2010 to 2019 as representing the push for common core standards and expanded teacher-based accountability as part of Race to the Top and NCLB waivers. Finally, we view the 2020 to present period as representing the post-COVID era.

pooled effects for programs serving between 400 and 999 students and 1,000 or more students have an average effect of 0.25 SD and 0.14 SD, respectively.

When we apply both sample restrictions to best approximate our target of inference, we arrive at our preferred set of estimates presented in Panel D of Table 5. These impacts range from 0.21 SD to 0.16 SD for U.S. tutoring programs evaluated using standardized test outcomes and operating at a scale of 400 to 999 and 1,000 or more students. There are four important points to highlight about this preferred set of estimates. First, they are between a third and a half as large as the pooled estimate using our full meta-analytic sample, suggesting that inferences made using the broader sample are not well-calibrated to tutoring programs at scale. Second, effect sizes between 0.16 SD and 0.21 SD are of medium to large magnitude and still very impressive for large-scale education interventions (Kraft, 2020). Third, our pooled effect size estimate for programs serving 1,000 students or more is very imprecisely estimated given the limited number of RCTs of tutoring programs at this scale that meet our target-of-inference-aligned inclusion criteria. Fourth, the wide prediction intervals associated with these estimates suggest that we should expect tutoring program effects to vary considerably, with some individual programs producing quite small or even negative effects and others resulting in sizable gains. We further test the robustness of this pattern of results by omitting estimates that are outliers and those from studies published since 2010, as shown in Panels E-G of Table 5 and Figure 3. This reduces the magnitude of the pooled effect size estimates among small-scale studies, but the overall pattern of declines at scale remains unchanged.

Evidence from large-scale non-experimental studies. Meta-analytic reviews of the literature on tutoring frequently restrict their focus to studies that employ RCTs in an effort to ensure researchers are identifying the unbiased, causal effect of tutoring. This restriction strengthens the internal validity of the pooled effect sizes, but may also limit the external validity of these findings (Tipton & Olsen, 2018). Large-scale RCTs are expensive and often require the active consent of participants, making them financially and logistically challenging to conduct. Our meta-analytic sample contains only nine studies that evaluate programs serving at least 1,000 students. This sparse data makes it difficult to accurately project plausible effects from tutoring programs taken to scale in larger U.S. school districts given a lack of common support.

We attempt to further inform our understanding of the plausible effects of tutoring by turning to studies of large-scale programs (n treated ≥ 1,000) that employ quasi-experimental methods. Much of the literature evaluating large-scale programs focuses on after-school tutoring provided by private tutoring organizations and funded by two federal initiatives, 21st Century Learning Centers and Supplemental Educational Services (SES) under the No Child Left Behind Act. Studies of these initiatives often evaluate programs across large districts and entire states with thousands of treated students and find effects that are notably smaller than those we find with our full meta-analytic sample (Deke et al., 2012; Heinrich et al., 2010, 2014; James-Burdumy et al., 2005; Ross et al., 2008; Springer et al., 2014; Zimmer et al., 2009, 2010). These small to medium effects (frequently ≤ 0.10 SD) may be fully explained by poor attendance at these off-site afterschool programs and their design features such as large student-tutor ratios and rotating tutors. However, the scale of the programs may also have contributed to their underwhelming results by influencing program design choices and implementation quality (see Kraft & Falken, 2020 for a fuller discussion).

Three recent studies from the post-COVID era provide more relevant assessments of ongoing attempts to integrate tutoring in the U.S. public school system at scale. Carbonari, Dewey, et al. (2024) evaluate the efforts of four mid- to large-sized districts to support students' academic recovery in math during the 2021-22 academic year by providing tutoring and additional instructional time. Using a value-added framework, they find estimates that are uniformly smaller than 0.04 SD and often precisely estimated null effects. These small impacts should be interpreted in the context of the first year of these large-scale initiatives when student attendance and staffing remained critical challenges for most districts. However, this same team of researchers have expanded their analyses to evaluate the effect of tutoring and small-group instruction across eight districts during the 2022-23 academic year and found similar results (Carbonari, DeArmond, et al., 2024). They document statistically insignificant estimates of the average effects of tutoring and small-group instruction of 0.03 SD in math and 0.07 SD in reading when pooling across tutoring programs that jointly served over 12,000 students.

Kraft et al. (2024) study efforts to scale tutoring in Metro-Nashville Public Schools (MNPS) over the course of two and half years to serve over 4,000 students by the spring of 2023. In contrast to the districts studied by Carbonari and colleagues, MNPS was largely successful at engaging students to attend tutoring frequently and staffing their program at scale by hiring their own teachers as tutors. Using an event study design, they find medium effects of tutoring on standardized tests scores in reading (0.09 SD), but no effects on test scores in math, on average.

Two recent evaluations of public tutoring programs implemented across the United Kingdom (U.K.) and in Victoria, Australia also provide early evidence of post-pandemic tutoring impacts at large scales in high-income contexts. Both analyses used matching methods that included baseline tests scores to reweight regression analyses, comparing the test score gains of tutored students to comparison-group students in the third year of these tutoring programs. Government Social Research, an evaluation agency within the U.K. Civil Service, found small to medium effects of the U.K. National Tutoring Programme on math (0.06 SD) and English achievement (0.03 SD) among Key Stage 2 students (years 3-6), but no effects on the achievement of Key Stage 4 students (years 10-11) in either subject (math -0.01 SD; English -0.00 SD) (Moore et al., 2024). The Victorian Auditor-General's Office found no significant effects of the state-wide Tutor Learning Initiative on students' achievement gains in math and reading across students in years 3 through 10 (Victorian Auditor-General's Office, 2024). Together, these quasi-experimental studies of large-scale tutoring programs are consistent with the overall pattern of declining tutoring impacts as program size increases.

Why do tutoring effects decline at scale?

The phenomenon of interventions becoming less effective when they are delivered to more students is a well-documented pattern in education research (Cheung & Slavin, 2016; Kraft, 2020, 2023). Understanding why this pattern exists for tutoring programs is critical to informing efforts to expand access to tutoring and maintain its effectiveness at scale. We posit and test four primary hypotheses that might explain this pattern.

Hypothesis #1: Declining effects do not reflect a true phenomenon but are instead due to selective reporting, standardization techniques, and/or spillover. It is possible that the negative relationship between program effects and program size is a product of the research process rather than a real pattern of differential effects. First, such a pattern could be caused by selective reporting that is more acute among studies with smaller samples. Here we define selective reporting as the phenomenon where studies that produce statistically insignificant results are less likely to result in academic publications. This could occur through multiple mechanisms including researchers being less likely to write papers when they find null results, researchers making subjective modeling decisions that push preferred estimates over traditional significance thresholds (i.e., *p*-hacking), and journals being less likely to publish studies that find null results (i.e., publication bias). Of course, researchers could also be systematically designing studies of programs that are likely to have larger effects to also have smaller sample sizes given less statistical power is necessary to detect larger effects.

We explore potential bias in three ways given that no single test can definitely rule out publication bias (McShane et al., 2016). First, we produce funnel plots and conduct a trim and fill analysis (Duval & Tweedie, 2000) to assess the degree of symmetry of our point estimates around the meta-analytic mean. An imbalance in publications falling on either side of vertical line at the center of the full plot would suggest potential bias, and lead the studies being imputed to make the data more symmetric. We do this at both the individual effect-size level and at the study level by collapsing multiple effects sizes to account for the nested nature of the data. As shown in Figure 4 and Appendix Table B1, we find no evidence of publication bias in our full sample of studies using this method. We then repeat these analyses after sub-setting our data into studies with fewer than 100 treated students versus at least 100 treated students and find no evidence of differential publication bias among small-sample studies.

Second, we test for evidence of *p*-hacking bias by plotting the *p*-values from our sample of effect sizes and examine whether there is an excess mass of *p*-values just below conventional significance thresholds in these distributions, following the intuition of Brodeur et al. (2020).[5] A visual inspection of Figure 5 reveals that the distribution of *p*-values is smooth across critical values for traditional significance thresholds in the full sample and in subsamples of smaller and larger sample studies. We then formally test for differential bunching below each conventional statistical threshold using a randomization test. This approach examines whether within a given

[5] We conduct these analyses using *p*-values rather than test statistics (i.e. *Z*-scores) because of the many small-sample studies in our review. This allows us to look at a sharp discontinuity for significance which would not be possible using the test statistics where threshold values change relative to the sample size.

window around a cut-point the *p*-values are binomial-distributed with equal probability. In Table 6, we show that we find little evidence to suggest there is differential bunching of estimates with *p*-values just below the 0.05 and 0.10 significance thresholds, nor any evidence that *p*-hacking is more common among small-scale studies. Only one of the six tests we run in our full sample using three different bandwidths for each threshold is marginally significant.

Our final test of selective reporting is to compare pooled effect sizes between peer-reviewed publications and other types of studies, such as working papers or reports, that have not gone through the peer-review process. If selective reporting was occurring because journals have been less likely to publish non-significant findings, we would expect to see larger average estimates from peer-reviewed than non-peer-reviewed studies. In Table B2, we show that this is indeed the pattern we find. Specifically, we observe an average pooled effect from studies in peer-reviewed journals of 0.45 SD versus 0.22 SD for non-peer-reviewed reports. We also check for differences by publication type across subsamples of studies with different sample sizes. For both peer-reviewed and non-peer-reviewed studies, we observe smaller effects for larger-sample studies. However, for studies with at least 1,000 treated students, the effects are quite a bit smaller for non-peer-reviewed papers (0.09 SD) than for peer-reviewed studies (0.33 SD).

We interpret these results with caution, especially given peer-reviewed status is correlated with other factors such as publication date. Our sample of non-peer-reviewed works skews more recent, and we know that more recent studies have demonstrated smaller pooled effects. These results are therefore not proof positive of selective reporting but are consistent with that possibility. In sum, we find mixed evidence on whether selective reporting could explain the pattern of declining pooled effects for programs implemented at a greater scale.

Publication bias is an area of active ongoing research and new methods for testing and addressing bias with clustered data may help us shed more light on this issue in the future.

A second possible statistical explanation for the differential pattern of effect sizes across smaller and larger tutoring programs is due to the standardization process. Tutoring programs typically target students that fall within a specific range of the performance distribution. We find that 94% of the studies we include in our meta-analysis describe some type of efforts to target students, with 89% of studies evaluating programs that specifically targeted low-performing students. As Fitzgerald and Tipton (2024) document, this targeting results in samples recruited to participate in RCTs to be more homogenous than the population as a whole. Targeted sampling reduces the variation in achievement among the study sample, artificially inflating the magnitude of the effect sizes when researchers standardize their outcome measure using sample-based estimates of its standard deviation. It is possible, if not likely, that the overall effect sizes from meta-analyses of tutoring are somewhat inflated because of this practice. This may also help to explain the pattern of attenuated effects we find if smaller-scale tutoring programs are able to more precisely target students, resulting in even more homogenous participant populations compared to larger-scale programs. Said another way, the pattern of declining effects by program size might be less pronounced if all studies had used an estimate of the SD of their test score outcome derived from nationally representative populations.

Finally, the presence of peer spillover effects could contribute to a differential pattern of tutoring effects by program size. A large body of evidence documents peer effects in K-12 education settings (Barrios-Fernandez, 2023). If being in the same class or school as a student receiving tutoring has positive spillover effects on non-tutored students *and* the magnitude of these effects increases with the concentration of treated students in a class or school, then larger-scale tutoring programs could differentially attenuate the treatment-control contrast and contribute to the pattern of declining effects we find. However, it is not obvious that this would happen in practice given that the concentration of treated students per class or school could be similar across smaller and larger programs if larger programs simply serve more schools.

Hypothesis #2: Scaling causes programs to systematically alter key design features. A second potential explanation for declining effects with scale is that leaders systematically change the design of tutoring programs for larger versus smaller scale interventions. To assess the evidence for this hypothesis, we first explore how key program features, which past research has highlighted as important elements of effective tutoring programs, change as programs are taken to scale. Table 7 reveals two systematic differences in program design features when comparing smaller verses larger programs. First, larger programs are substantially less likely to tutor students individually. Programs serving over 400 students are roughly 10 percentage points less likely to rely on 1:1 student-tutor ratios than small programs that serve fewer than 100 students. Second, larger programs tend to deliver less dosage, primarily by shortening the number of weeks tutoring programs run. Here the relationship is not entirely monotonic, with the smallest tutoring programs offering moderate dosage, middle-sized tutoring programs with the highest total dosage and larger tutoring programs offering the least. For example, on average, programs that serve 100 to 399 students delivered 39 total tutoring hours while those serving greater than 1,000 delivered 27 total hours. Unexpectedly, we see that larger programs are even slightly more likely to use teachers and paraprofessionals as tutors and to provide a high degree of supervision and support to tutors – characteristics hypothesized to promote larger effects.

As another test, we examine whether the negative relationship between program effects and size is attenuated when we control for the full range of observable program characteristics in a meta-regression framework. We do this by comparing the results of two meta-regressions. The first model shown in Table 8 reports coefficients from binned sample size indicators which capture the clear negative relationship relative to the omitted category of studies evaluating small programs serving fewer than 100 students. We then add a large set of control variables including indicators for the intersection of schooling-level and subject, researcher generated assessments, studies from international contexts, publication decade, and our full set of tutoring program characteristics. We find that the strength of the negative relationship between program size and effects is reduced by roughly 30%, suggesting changing program characteristics do account for a portion of the overall negative relationship between program size and effects on achievement.

Notably, only a few study features and program characteristics appear systematically related to effect sizes when included in our fully controlled meta-analytic model. In addition to sample size, researcher-generated tests produce meaningfully larger effect sizes relative to third-party standardized tests (0.22 SD), tutoring outside of the school day has a strong negative association with effect sizes relative to during the school day (-0.19 SD), and an indicator for studies that did not report a specific tutor-student ratio – perhaps suggestive of larger and flexible ratios – also has a strong negative association compared to those with a 1:1 ratio (-0.19 SD). We similarly find a negative association with using a specified tutor type not in our major categories, relative to a teacher (-0.16 SD), although this group mostly consists of community members and/or volunteers. While we find significant positive associations with two bins of total dosage hours relative to receiving at least 60 hours of treatment, these results don't show a monotonic relationship between dosage and impacts, suggesting more systematic exploration is needed.

Hypothesis #3: Heterogeneous tutoring effects cause the marginal student to benefit less as tutoring programs expand. The attenuation of tutoring effects as program sizes increase may also be a product of the heterogeneous effects of tutoring across students. Prior research has found that tutoring may be more effective for students who are lower-performing prior to tutoring (Kraft, 2015; Robinson et al., 2024), Black students (Fryer & Howard-Noveck, 2020) and students from low-income families (Carlana & La Ferrara, 2024). It is plausible that smaller-scale tutoring programs appear more effective because they are better able to target students who stand to benefit the most, on average, from the programming. As tutoring programs scale, they may be expanding to serve students who will benefit less, on average.

We explore this by comparing weighted averages of student characteristics in our sample of RCTs, disaggregating by the size of the tutoring program, in Table 9. This comparison reveals a clear pattern where studies of smaller tutoring programs serve larger percentages of historically marginalized students. Students in smaller tutoring programs were 13 percentage points more likely to be English learners, 11 percentage points more likely to be receiving special education services, 8 percentage points more likely to be from low-income backgrounds, 4 percentage points more likely to be Hispanic, and 2 percentage points more likely to be Native American. These sizable differences in the characteristics of students served by smaller and larger tutoring programs are likely to attenuate the estimated effects of tutoring as program scale increases.

A related possibility that we cannot directly test with our data is that smaller programs treat student populations that are more homogenous. Homogeneity may make implementation easier because there is less of a need to tailor interventions to a variety of student achievement levels or other unique needs. Expanding tutoring programs might mean programming is provided to a more diverse group of students with a wider set of challenges, making it more difficult to produce large impacts among increasingly heterogenous groups.

Hypothesis #4: Implementation quality declines as tutoring programs scale. A final hypothesis for why we observe smaller impacts for larger programs is that the quality of program implementation declines as tutoring programs are brought to scale. Imagine, for example, two tutoring programs with the exact same intended program design features (e.g., high-dosage, 1:1 ratios, paraprofessional tutors), but one serves a small number of students at a single school and the other is brought to scale districtwide. Administrative needs are likely higher for the large-scale program. Small programs may be more likely to represent pilot efforts led by uniquely trail-blazing, motivated, and talented leaders, whereas administrators recruited to run large programs may not be as effective, on average. Implementation quality could also suffer if the effectiveness of the average tutor is lower for larger programs than for smaller programs. However, if tutoring screening tools are only weakly related to tutor performance, then tutor quality may not decline with scale (Davis et al., 2017). It may be more challenging to coordinate communication between tutors and teachers for large-scale programs. There may simply be less oversight with a greater number of tutoring sites, making it more difficult to ensure fidelity of implementation to program models for larger interventions. It may be that even if the intended program design remains constant as programs expand, the delivered dosage drops if student attendance suffers or time-on-task declines as programs expand. This would still be consistent with our pattern of results, given that our coding reflects intended measures of dosage rather than the actual number of total hours of tutoring treated students received.

Unfortunately, most tutoring RCTs do not directly measure implementation quality, limiting what we can say about this hypothesis using our meta-analytic dataset. However, survey data and several recent studies on post-COVID tutoring efforts do point to significant implementation challenges. To start, the majority of K-12 public school principals report experiencing barriers (e.g., funding, timing, or staffing challenges) that limited their ability to effectively provide tutoring on the nationally representative SPP survey (National Center for Education Statistics, 2024). The aforementioned "Road to Recovery" (R2R) evaluation of large-scale academic recovery efforts by Carbonari, Dewey, et al. (2024) documents how districts fell well short of leaders' intended expectations with regard to both the number of students served and the intensity of the interventions. Interviews with district leaders revealed challenges with engaging targeted students. This is consistent with SPP survey results showing that, among schools that provided tutoring, larger schools had somewhat lower student participation rates (National Center for Education Statistics, 2024). Research on an opt-in virtual tutoring program in the spring of 2021 further illustrates challenges related to low student participation, especially among struggling students who might benefit most from tutoring (Robinson et al., 2022).

Buy-in is also a problem identified amongst staff. Programs that appeared to successfully scale high-quality tutoring after the pandemic emphasized the importance of district-level leadership, goal setting, buy-in from school leaders and teachers, a willingness to rethink scheduling, the pursuit of multiple funding sources, and the ability to make difficult choices about spending trade-offs (Cohen, 2024). Leaders in the R2R districts highlighted staffing challenges related to pandemic surges, a tight labor market, and limited district capacity for recruitment and human resources management. These issues of staffing challenges and organizational capacity are echoed by findings from a qualitative study on programs in two urban districts (Makori et al., 2024). These implementation challenges do not appear to be solely a function of acute post-pandemic conditions, as the R2R team's follow up report examining academic recovery efforts during 2022-23 revealed that difficulties persisted (Carbonari, DeArmond, et al., 2024). Across the majority of interventions examined by the R2R study, including tutoring, fewer students participated and for less time than intended. Consistent with our meta-analytic findings, the rare R2R tutoring programs that generated positive impacts on test scores were those implemented on a small scale. Finally, leaders interviewed for the R2R report also pointed to their need to adapt tutoring program designs—sometimes departing from best practices—to align with federal, state, and local policies. This is likely to remain a challenge as schools and districts look to a range of federal, state, and local funding sources to support tutoring programs after the COVID-relief funding runs dry (Accelerate, 2023; Cohen, 2024).

How might policymakers approach the challenge of maintaining effectiveness at scale?

Like so many complex interventions, the efficacy of tutoring programs may lie in the combination of program design features rather than any single characteristic. Prior literature has focused on a bundle of program features that research suggests are associated with larger effects, aligned with what is sometimes described as "high-quality," "high-dosage," or "high-impact" programs (e.g., Robinson et al., 2021). This bundle of features includes in-person programming, delivered at school during school hours, with a student-tutor ratio of no more than 3:1, meeting at least 3 times per week, ensuring a high overall dosage (which we proxy for with at least 15 hours of total tutoring), and using a provided curriculum.[6]

When we test whether the combination of these features are greater than the sum of their parts, we find encouraging results. Specifically, the overall pattern of declining effect sizes persists among tutoring programs that utilize a bundled package of recommended design features but the attenuation at scale is less pronounced. When we isolate only individual features of this bundle, effects continue to erode to varying degrees as programs scale (Appendix Table B3).

[6] Sustained relationships between tutors and students and data-informed instruction based on formative assessments are also recommended. Unfortunately, most studies did not provide information to allow us to code these features.

That said, this erosion is milder for in-person programs providing at least 15 hours of tutoring. As shown in Figure 6, while the pooled effect among studies of programs serving between 100-399 students declines by 42% relative to programs serving 99 students or fewer in the full sample, it only declines by 18% in the restricted sample of studies with the bundled package of design features. The decline among programs serving 400-999 students is also slightly less pronounced, dropping 54% in the full sample and 44% in the bundled package sample. Perhaps most striking is that when we restrict our analysis to studies evaluating U.S. programs based on standardized test measures published after 2009 and omit outliers, we see no attenuation of program effects across sample size, at least for programs serving fewer than 1,000 students.[7]

What does the research suggest about modifying program design features to reduce costs and increase scalability?

Although the bundled package of program features appears to help sustain program effectiveness at scale, several aspects are costly and can be difficult to implement at scale. Here we explore the potential implications of modifying specific program features.

Moving Tutoring Online: Many districts and programs have adopted online tutoring to access a larger potential supply of tutors. How might this affect the efficacy of tutoring? When we limit our sample to the 59 estimates of tutoring delivered virtually (drawn from 6 unique studies), the pooled estimate reported in Table 11a is 0.07 SD.[8] This is substantially smaller than the unadjusted pooled estimate of in-person program impacts of 0.44 SD and even substantially smaller than our preferred pooled estimates of expected impacts for our target of inference, 0.16

[7] We are unable to estimate this number for programs serving more than 1,000 students because only one of the larger scale programs evaluated in our sample implemented the full bundle of recommended program characteristics.
[8] We observe the following virtual tutoring studies: Fesler et al. (2023), Gortazar et al. (2023), Kraft et al. (2022), Loeb et al. (2023), Roschelle et al. (2020), and Torgerson et al. (2016). We exclude Carlana and La Ferrara (2021) because their achievement outcome pools across math, Italian, and English.

to 0.21 SD (Table 5). However, results from our meta-analytic regressions presented in Table 8 suggest these smaller effects are likely driven by other program features. Conditional on our extensive set of codes for observable program features, we estimate a positive but statistically insignificant coefficient when comparing virtual tutoring programs to those in person.

Unfortunately, there are no studies to our knowledge that provide a direct causal comparison between virtual and in-person programs, which would be a major contribution to the field. Another limitation of the literature is that most studies of virtual tutoring in our sample were conducted in the post-COVID era, and some during the early part of the pandemic when challenges were most acute. These studies may not generalize well to non-pandemic times or to contexts that differ in important ways from those where leaders were willing to partner with researchers amid pandemic recovery. Two new studies of virtual tutoring released in 2024 find estimates that are similar to (Ready et al., 2024) or somewhat larger than (Carlana & La Ferrara, 2024) the magnitude of our pooled estimates for virtual programs. We read this evidence as suggestive that online tutoring has the potential to be an effective approach to addressing scaling challenges when accompanied by other effective program design characteristics.

Increasing Student-Tutor Ratios: The cost of tutoring is driven largely by tutor compensation. Many districts and tutoring organizations have chosen to increase student-tutor ratios as a means of expanding access while managing costs. In Table 11a, we report pooled effect estimates by student-tutor ratio and find somewhat larger impacts, on average, for programs with lower ratios. Using the full sample, we estimate pooled effects of 0.43, 0.41, 0.30, and 0.34 SD for 1:1, 2:1, 3:1, and 4:1 programs, respectively. The effects for programs with 5 or more students per tutor are substantially larger with the full sample (0.91 SD), but this result is not robust to excluding studies outside of our target of inference. The overall pattern of declining effects persists when we focus on U.S. tutoring programs evaluated using standardized tests and when we further restrict to those studies for which we have more confidence in the quality of research methods. Importantly, even programs with ratios of 5 or more generate meaningful impacts, on average, between 0.24 SD and 0.29 SD. When we examine student-tutor ratios in a meta-regression framework, we find a pattern of larger effects for smaller student-tutor ratios, although none of the point estimates are statistically significant.

Evidence from the ten studies that experimentally vary student-tutor ratios, summarized in Table 12a, provide a range of contrasts from 1:1 versus 2:1 ratios (Carlana & La Ferrara, 2024; Loeb et al., 2023; Vadasy & Sanders, 2008) to 4:1 to 13:1 ratios (Vaughn et al., 2010). Most examine interventions with elementary students (Clarke et al., 2017, 2020, 2023; Doabler et al., 2019; Loeb et al., 2023; Schwartz et al., 2012; Vadasy & Sanders, 2008) except for three with middle schoolers (Carlana & La Ferrara, 2024; Kraft & Lovison, 2024; Vaughn et al., 2010). The effect size differences most often favor smaller ratios but are not always large in magnitude and do not typically achieve statistical significance. However, many of these studies are underpowered to detect small differences in effects between treatment arms. In short, the existing research suggests that lower ratios produce larger effects, but it is possible to deliver tutoring in pairs or small groups and maintain meaningful effects.

Using Peer Tutors: An alternative approach to scaling tutoring on a fixed budget is to enlist K-12 students as peer tutors. We find that pooled effect sizes for peer tutoring in our full sample are an impressive 0.36 SD, shown in Table 11b. Our meta-analytic regression (Table 8) suggests peer tutoring is equally as effective as tutoring by teachers, conditional on other program and study characteristics, with a non-significant difference of just 0.06 SD in favor of teachers relative to peer tutors. We know of only one study that randomizes students to different tutor types. Mathes et al., (2003) uses a partially matched and partially randomized design to compare teachers who implemented small group (4-5:1) instruction verses overseeing pairs of students who used Peer Assisted Learning Strategies (PALS). They find effect sizes of 0.70 SD for teacher directed small-group instruction and 0.55 SD for peer-assisted instruction. We see the limited evidence for scaling with peer tutoring as encouraging but incomplete.

Decreasing Dosage: A fourth common approach to scaling tutoring while controlling costs is to reduce overall dosage. Pooled effect estimates presented in Table 11b do not reveal a clear monotonic trend between dosage hours and program impacts, particularly in our more restricted samples. Across the full sample as well as the more restricted, generalizable samples, programs offering over 60 hours of tutoring consistently have the smallest impacts. In our preferred policy-relevant subsample, the greatest magnitude of effect is for programs providing 15-29 hours of tutoring (0.41 SD). However, these pooled estimates may be inflated or deflated if other study characteristics are correlated with certain amounts of dosage. When employing meta-regression to control for a variety of program features and study characteristics, we do not find consistent statistically significant differences based on the total hours of dosage, as shown in Table 8. If anything, the pattern of results is counter-intuitive with programs offering higher dosage showing smaller effects. Here selection bias still poses a potential challenge if, for example, more intensive programs target particularly struggling students.

Evidence from four studies that randomly assigned students to different doses of tutoring to isolate the causal impact of dosage suggests some benefits of greater dosage. As shown in Table 12b, three of these studies evaluate elementary school programs (Al Otaiba et al., 2005; Begeny, 2011; Wanzek & Vaughn, 2008) and one middle school program (Carlana & La Ferrara, 2021). These studies provide a range of contrasts, for example, comparing four versus nine total hours of tutoring (Begeny, 2011) to comparing 36 hours versus 72 hours (Al Otaiba et al., 2005). More often than not, these studies show greater effect sizes for programs providing higher than lower dosages. In short, studies that experimentally vary dosage suggest that decreasing dosage may attenuate effects. However, we still have much to learn about dosage effects related to days per week, hours per session, hours per week, total weeks, and whether the ideal design varies for different age groups, subjects, or by other factors.

Discussion

Evidence-based policymaking has increasingly become the standard in education, particularly as practitioners look to implement proven approaches to accelerate students' academic growth after the substantial disruptions caused by the COVID-19 pandemic. While this trend is encouraging, it places increased importance on the external validity of research. Even well designed and implemented RCTs – the gold standard approach to causal inference – offer incomplete information to policymakers and practitioners if the evidence they produce is at arm's length from the realities of implementing education policies and practice at scale. Meta-analyses that pool evidence across multiple studies seemingly offer evidence with strong external validity, but aggregating across multiple studies with limited generalizability does not make the results valid for a very different target of inference.

Our study illustrates the importance of carefully considering the alignment between the research evidence and the policy target of inference. We find that attempts to better harmonize our meta-analytic sample of 265 studies to the target of inference used by most policymakers – large-scale tutoring programs in the U.S. aiming to increase student performance on standardized tests – substantially reduces the pooled effect sizes we find. This attenuation is largely driven by the declining impacts of tutoring programs as they scale – a pattern that is common across the education research literature.

This pattern of declining effects at scale often leads to a circular argument that, “the program works when implemented with fidelity, it just wasn’t implemented correctly when taken to scale.” Alternatively, one might ask, “If implementation becomes systematically more difficult at scale, then does a program really work?” We see four possible responses to this challenge: 1) start small, learn, iterate, and engage in the hard but critical work to scale vertically (i.e. expanding program size) over time while maintaining program fidelity, 2) redesign the program to be easier to implement at scale, 3) adopt a more flexible approach to scaling that allows for localized adaptation, and/or 4) decide that a program is best delivered in a small-scale format and focus on horizontal scaling (replicating small programs).

To be clear, we view the more target-equivalent estimates of the effects of tutoring we find as still meaningful and policy-relevant (Kraft, 2020). We continue to see tutoring as one of the most promising evidence-based approaches to accelerating student achievement. If districts could leverage tutoring at scale for those students whose learning was most negatively affected by the pandemic and produce effects similar to our policy-relevant estimates, it would be a huge success. In fact, several recent experimental studies of tutoring programs implemented post-COVID at a medium scale (Carlana & La Ferrara, 2024; Cortes et al., 2024; Gortazar et al., 2024) and at a large scale (Robinson et al., 2024) find effects on par with those from our target-aligned pooled effect sizes.

That said, we also think it is equally important for policymakers and practitioners alike to have more grounded expectations about what tutoring can accomplish. Several other recent studies using both experimental and quasi-experimental methods suggest early attempts to scale tutoring in the U.S. have produced quite small effects (Carbonari, DeArmond, et al., 2024; Carbonari, Dewey, et al., 2024; Kraft et al., 2024; Ready et al., 2024). Outsized expectations can lead policymakers and practitioners to become disillusioned when they fail to realize the eye-popping effect sizes of small-scale, boutique tutoring programs implemented under favorable circumstances among students who often opt into participating, particularly when meta-analytic estimates mask those contextual factors. Unrealistic expectations can also lead policymakers to mistakenly rely on a single or limited set of interventions when multiple interrelated programs may be needed to achieve their goals. Contextualizing tutoring program effects relative to their costs will also be critical for identifying sustainable models (Kohlmoos & Steinberg, 2024).

New technology may also present opportunities to scale tutoring with greater fidelity while maintaining program effects and reducing per-pupil costs. Recent studies suggest that computer-assisted learning programs paired with tutoring (Bhatt et al., 2024) or integrated into core academic classes (Oreopoulos et al., 2024) can support effective instruction, potentially reducing common obstacles to scaling tutoring. There is growing interest in the potential of generative artificial intelligence to offer effective tutoring at scale, although early programs appear to fall well short of this goal (Barnum, 2024). We remain optimistic about the potential of these new technologies but emphasize that the benefits of human tutoring likely extend far beyond student performance on standardized tests, to say nothing about the value of tutoring for the tutor. Human tutoring offers the opportunity for authentic personal connections and social interactions that can contribute to student development; it also creates volunteer and employment opportunities and valuable experiences for those interested in pursuing a career in education.

Conclusion

Efforts to integrate tutoring at scale into the U.S. K-12 public education system are at a critical juncture. New evidence documenting the mixed results of early efforts to expand access to tutoring during the 2021-22 and 2022-23 academic years is emerging right as large-scale federal funding to support tutoring is ending. With this paper, we aim to inform ongoing efforts to refine tutoring programs when implemented at scale and better calibrate expectations for what these programs are capable of accomplishing. Our findings highlight the importance of conducting research that considers both internal and external validity to best inform policy and practice.

Our analyses suggest that a bundled package of program features hypothesized to promote effective tutoring does guard against some of the attenuation that occurs as programs expand. It remains an open question whether adapting individual features of this bundle – such as moving tutoring online, increasing student-tutor ratios, using peer tutors, or decreasing dosage – can be done without compromising effectiveness. Such changes may attenuate effects but still be an equally, if not more, cost-effective way to deliver tutoring at scale. Our hope is that as policymakers experiment with new tutoring models, they will partner with researchers to learn about the impacts of these adaptations. Continued efforts to integrate individualized instruction into the U.S. K-12 education system would benefit from a decades-long approach that focuses first on establishing effectiveness and then on scaling, rather than the other way around.

References

Accelerate. (2023). *Beyond recovery: Funding high-impact tutoring for the long term*. Accelerate. https://accelerate.us/beyond-recovery/

Al Otaiba, S., Schatschneider, C., & Silverman, E. (2005). Tutor-assisted intensive learning strategies in kindergarten: How much is enough? *Exceptionality*, *13*(4), 195–208.

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research*, *90*(1), 6–23.

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy*, *20*(2), 198–212. https://doi.org/10.1093/oxrep/grh011

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, *1*(Volume 1, 2009), 151–178.

Barnum, M. (2018). Why "personalized learning" advocates like Mark Zuckerberg keep citing a 1984 study—And why it might not say much about schools today. *Chalkbeat*.

Barnum, M. (2024). We tested an ai tutor for kids. It struggled with basic math. *The Wall Street Journal*.

Barrios-Fernandez, A. (2023). Peer Effects in Education. In *Oxford Research Encyclopedia of Economics and Finance*.

Begeny, J. C. (2011). Effects of the Helping Early Literacy with Practice Strategies (helps) reading fluency program when implemented at different frequencies. *School Psychology Review*, *40*(1), 149–157. https://doi.org/10.1080/02796015.2011.12087734

Bhatt, M., Guryan, J., Khan, S., LaForest, M., & Bhavya Mishra. (2024). Can technology facilitate scale? Evidence from a randomized evaluation of high dosage tutoring. *NBER Working Paper*, *32510*. https://doi.org/10.17605/OSF.IO/UW8EH

Borenstein, M., Higgins, J., Hedges, L., & Rothstein, H. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18.

Boveda, M., Ford, K. S., Frankenberg, E., & López, F. (2023). Editorial vision 2022–2025. *Review of Educational Research*, *93*(5), 635–640.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *The American Economic Review*, *110*(11), 3634–3660.

Callen, I., Goldhaber, D., Kane, T. J., McDonald, A., McEachin, A., & Morton, E. (2024). *Pandemic learning loss by student baseline achievement: Extent and sources of heterogeneity* (292–0224; CALDER Working Paper). American Institutes for Research.

Carbonari, M. V., DeArmond, M., Dewey, D., Dizon-Ross, E., & Goldhaber, D. (2024). *Impacts of academic recovery interventions on student achievement in 2022-23* (303–0724; CALDER Working Paper). American Institutes for Research.

Carbonari, M. V., Dewey, D., Kane, T. J., Muroga, A., DeArmond, M., Dizon-Ross, E., Goldhaber, D., Morton, E., Davison, M., & Hashim, A. K. (2024). *The impact and implementation of academic interventions during Covid: Evidence from the road to recovery project* (CALDER Working Paper). American Institutes for Research.

Carlana, M., & La Ferrara, E. (2021). *Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic* (21–350; EdWorkingPaper).

Carlana, M., & La Ferrara, E. (2024). *Apart but connected: Online tutoring, cognitive outcomes, and soft skills* (Working Paper 32272). National Bureau of Economic Research.

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

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), 233285841770689. https://doi.org/10.1177/2332858417706899
Clarke, B., Doabler, C. T., Sutherland, M., Kosty, D., Turtura, J., & Smolkowski, K. (2023). Examining the impact of a first grade whole number intervention by group size. *Journal of Research on Educational Effectiveness*, *16*(2), 326–349.
Clarke, B., Doabler, C. T., Turtura, J., Smolkowski, K., Kosty, D. B., Sutherland, M., Kurtz-Nelson, E., Fien, H., & Baker, S. K. (2020). Examining the efficacy of a kindergarten mathematics intervention by group size and initial skill: Implications for practice and policy. *The Elementary School Journal*, *121*(1), 125–153. https://doi.org/10.1086/710041
Cohen, L. (2024). *Learning curve: Lessons from the tutoring revolution in public education*. FutureEd & the National Student Support Accelerator.
Cortes, K., Kortecamp, K., Loeb, S., & Robinson, C. (2024). *A scalable approach to high-impact tutoring for young readers: Results of a randomized controlled trial* (Working Paper 32039). National Bureau of Economic Research. https://doi.org/10.3386/w32039
Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A., & Steingrimsson, J. A. (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, *31*(3), 334.
Davis, J. M. V., Guryan, J., Hallberg, K., & Ludwig, J. (2017). *The economics of scale-up* (w23925). National Bureau of Economic Research. https://doi.org/10.3386/w23925
Deke, J., Dragoset, L., Bogen, K., & Gill, B. (2012). *Impacts of Title I Supplemental Educational Services on student achievement* (2012–4053). National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED532016
Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*(2), 243–282.
DiMarco, B., & Jordan, P. W. (2022, February 24). What will $50b in COVID school relief funding buy? *The74Million*.
Doabler, C. T., Clarke, B., Kosty, D., Kurtz-Nelson, E., Fien, H., Smolkowski, K., & Baker, S. K. (2019). Examining the impact of group size on the treatment intensity of a tier 2 mathematics intervention within a systematic framework of replication. *Journal of Learning Disabilities*, *52*(2), 168–180. https://doi.org/10.1177/0022219418789376
Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89–98. https://doi.org/10.1080/01621459.2000.10473905
Esterling, K., Brady, D., & Schwitzgebel, E. (2024). *The necessity of construct and external validity for deductive causal inference*. https://doi.org/10.31219/osf.io/2s8w5
Fesler, L., Gu, A., & Chojnacki, G. (2023). Air tutors' online tutoring: Math knowledge impacts and participant math perceptions. Mathematica. https://eric.ed.gov/?id=ED628638
Fitzgerald, K. G., & Tipton, E. (2024). Using extant data to improve estimation of the standardized mean difference. *Journal of Educational and Behavioral Statistics*.
Fryer, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experimentsa. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Vol. 2, pp. 95–322). North-Holland.
Fryer, R. G., & Howard-Noveck, M. (2020). High-Dosage Tutoring and Reading Achievement: Evidence from New York City. *Journal of Labor Economics*, *38*(2), 421–452.

Goldhaber, D., Kane, T. J., McEachin, A., & Morton, E. (2022, November 16). Opinion | To help students shoot for the moon, we must think bigger and bolder. *Washington Post*.

Goldhaber, D. & Falken, G. (2024). ESSER and student achievement: Assessing the impacts of the largest one-time federal investment in K12 schools. CALDER Working Paper No. 301-0624.

Gortazar, L., Hupkau, C., & Roldán-Monés, A. (2024). Online tutoring works: Experimental evidence from a program with vulnerable children. *Journal of Public Economics*, *232*.

Gortazar, Lucas, Hupkau, Claudia, & Roldan, Antonio. (2023). *Online tutoring works: Experimental evidence from a program with vulnerable children* (EdWorkingPaper 23–743). Annenberg Institute for School Reform at Brown University.

Hedberg, E. C. (2014). *ROBUMETA: Stata module to perform robust variance estimation in meta-regression with dependent effect size estimates* [Stata]. Boston College Department of Economics. https://ideas.repec.org//c/boc/bocode/s457219.html

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hedges, L., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

Heinrich, C. J., Burch, P., Good, A., Acosta, R., Cheng, H., Dillender, M., Kirshbaum, C., Nisar, H., & Stewart, M. (2014). Improving the implementation and effectiveness of out-of-school-time tutoring. *Journal of Policy Analysis and Management*, *33*(2), 471–494.

Heinrich, C. J., Meyer, R. H., & Whitten, G. (2010). Supplemental education services under No Child Left Behind: Who signs up, and what do they gain? *Educational Evaluation and Policy Analysis*, *32*(2), 273–298. https://doi.org/10.3102/0162373710361640

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, *2*(3), 172–177.

Inns, A. J., Lake, C., Pellegrini, M., & Slavin, R. (2019). *A quantitative synthesis of research on programs for struggling readers in elementary schools* (Best Evidence Encylopedia). Center for Research and Reform in Education.

James-Burdumy, S., Dynarski, M., Moore, M., Deke, J., Mansfield, W., Pistorino, C., & Warner, E. (2005). When schools stay open late: The national evaluation of the 21st Century Community Learning Centers program. Final report. U.S. Department of Education.

Kohlmoos, L., & Steinberg, M. P. (2024). *Contextualizing the impact of tutoring on student learning: Efficiency, cost effectiveness, and the known unknowns*. Accelerate.

Kraft, M. A. (2015). How to Make Additional Time Matter: Integrating Individualized Tutorials into an Extended Day. *Education Finance and Policy*, *10*(1), 81–116.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kraft, M. A. (2023). The effect-size benchmark that matters most: Education interventions often fail. *Educational Researcher*, *52*(3), 183–187.

Kraft, M. A., Edwards, D. S., & Cannata, M. (2024). Scaling tutoring in a large urban school district: Program dynamics and causal effects. *NBER Working Paper*.

Kraft, M. A., & Falken, G. T. (2021). A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*, *7*, 233285842110428.

Kraft, M. A., List, J. A., Livingston, J. A., & Sadoff, S. (2022). Online tutoring by college volunteers: Experimental evidence from a pilot program. *AEA Papers and Proceedings*, *112*, 614–618. https://doi.org/10.1257/pandp.20221038

Kraft, M. A., & Lovison, V. S. (2024). The effect of student-tutor ratios: Experimental evidence from a pilot online math tutoring program. In *EdWorkingPapers.com* (EdWorkingPaper). Annenberg Institute at Brown University. https://edworkingpapers.com/ai24-976

Littell, J. H. (2024). The logic of generalization from systematic reviews and meta-analyses of impact evaluations. *Evaluation Review*, *48*(3), 427–460.

Loeb, S., Novicoff, S., Pollard, C., Robinson, C., & White, S. (2023). *The effects of virtual tutoring on young readers: Results from a randomized controlled trial*. National Student Support Accelerator.

Makori, A., Burch, P., & Loeb, S. (2024). *Scaling high-impact tutoring: School level perspectives on implementation challenges and strategies* (24–923; EdWorkingPaper).

Mathes, P. G., Torgesen, J. K., Clancy-Menchetti, J., Santi, K., Nicholas, K., Robinson, C., & Grek, M. (2003). A comparison of teacher-directed versus peer-assisted instruction to struggling first-grade readers. *The Elementary School Journal*, *103*(5), 459–479.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Moore, E., Morton, C., Schwendel, G., & Welbourne, S. (2024). *National Tutoring Programme year 3: Impact report*. Department for Education, Government Social Research.

National Center for Education Statistics. (2024). *School Pulse Panel: Responses to the pandemic and efforts toward recovery*. U.S. Dept. of Education, Institute of Education Sciences.

National Student Support Accelerator. (2023). *A snapshot of state tutoring policies*. https://studentsupportaccelerator.org/briefs/snapshot-state-tutoring-policies

Nickow, A., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence* (Working Paper 27476). National Bureau of Economic Research. https://doi.org/10.3386/w27476

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. https://doi.org/10.3102/00028312231208687

Oreopoulos, P., Gibbs, C., Jensen, M., & Price, J. (2024). Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence. *NBER Working Paper*, *32388*.

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. *AERA Open*, *7*, 2332858420986211.

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46.

Polanin, J., Zhang, Q., Taylor, J., Williams, R., Joshi, M., & Burr, L. (2023). Evidence gap maps in education research. *Journal of Research on Educational Effectiveness*, *16*(3), 532–552.

Pritchett, L., & Sandefur, J. (2015). Learning from experiments when context matters. *American Economic Review*, *105*(5), 471–475. https://doi.org/10.1257/aer.p20151016

Ready, D. D., McCormick, S. G., & Shmoys, R. J. (2024). *The effects of in-school virtual tutoring on student reading development: Evidence from a short-cycle randomized controlled trial* (24–942; EdWorkingPaper). Annenberg Institute at Brown University.

Robinson, C. D., Bisht, B., & Loeb, S. (2022). The inequity of opt-in educational resources and an intervention to increase equitable access. *Annenberg EdWorkingPaper*, *22–654*.

Robinson, C. D., Kraft, M. A., Loeb, S., & Schueler, B. E. (2021). *Accelerating student learning with high-dosage tutoring.* EdResearch for Recovery Project.

Robinson, C. D., Pollard, C., Novicoff, S., White, S., & Loeb, S. (2024). The effects of virtual tutoring on young readers: Results from a randomized controlled trial. *EdWorkingPapers No. 24-955.*

Roschelle, J., Cheng, B. H., Hodkowski, N., Neisler, J., & Haldar, L. (2020). Evaluation of an online tutoring program in elementary mathematics. Digital Promise.

Ross, S., Potter, A., Paek, J., McKay, D., Sanders, W., & Ashton, J. (2008). Implementation and outcomes of Supplemental Educational Services: The Tennessee state-wide evaluation study. *Journal of Education for Students Placed at Risk (JESPAR)*, *13*(1), 26–58.

Schwartz, R. M., Schmitt, M. C., & Lose, M. K. (2012). Effects of teacher-student ratio in response to intervention approaches. *Elementary School Journal*, *112*(4), 547–567.

Slough, T., & Tyson, S. A. (2023). External validity and meta-analysis. *American Journal of Political Science*, *67*(2), 440–455. https://doi.org/10.1111/ajps.12742

Springer, M. G., Pepper, M. J., & Ghosh-Dastidar, B. (2014). Supplemental Educational Services and student test score gains: Evidence from a large, urban school district. *Journal of Education Finance*, *39*(4), 370–403.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes. *Research Synthesis Methods*, *5*(1), 13–30. https://doi.org/10.1002/jrsm.1091

Tipton, E., Bryan, C., Murray, J., McDaniel, M. A., Schneider, B., & Yeager, D. S. (2023). Why meta-analyses of growth mindset and other interventions should follow best practices for examining heterogeneity: Commentary on Macnamara and Burgoyne (2023) and Burnette et al. (2023). *Psychological Bulletin*, *149*(3–4), 229–241.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524.

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., Jefferson, L., Mitchell, N., & Torgerson, D. (2016). Affordable online maths tuition: Evaluation report and executive summary. Education Endowment Foundation.

Tyack, D. B. (1974). *The one best system: A history of American urban education*. Harvard University Press.

Vadasy, P. F., & Sanders, E. A. (2008). Code-oriented instruction for kindergarten students at risk for reading difficulties. *Reading and Writing*, *21*(9), 929–963.

Vaughn, S., Wanzek, J., Wexler, J., Barth, A., Cirino, P. T., Fletcher, J., Romain, M., Denton, C. A., Roberts, G., & Francis, D. (2010). The relative effects of group size on reading progress of older students with reading difficulties. *Reading and Writing*, *23*(8), 931–956.

Victorian Auditor-General's Office. (2024). *Effectiveness of the Tutor Learning Initiative* Independent assurance report to Parliament.

von Hippel, P. T. (2024, March 7). Two-sigma tutoring: Separating science fiction from science fact. *Education Next*.

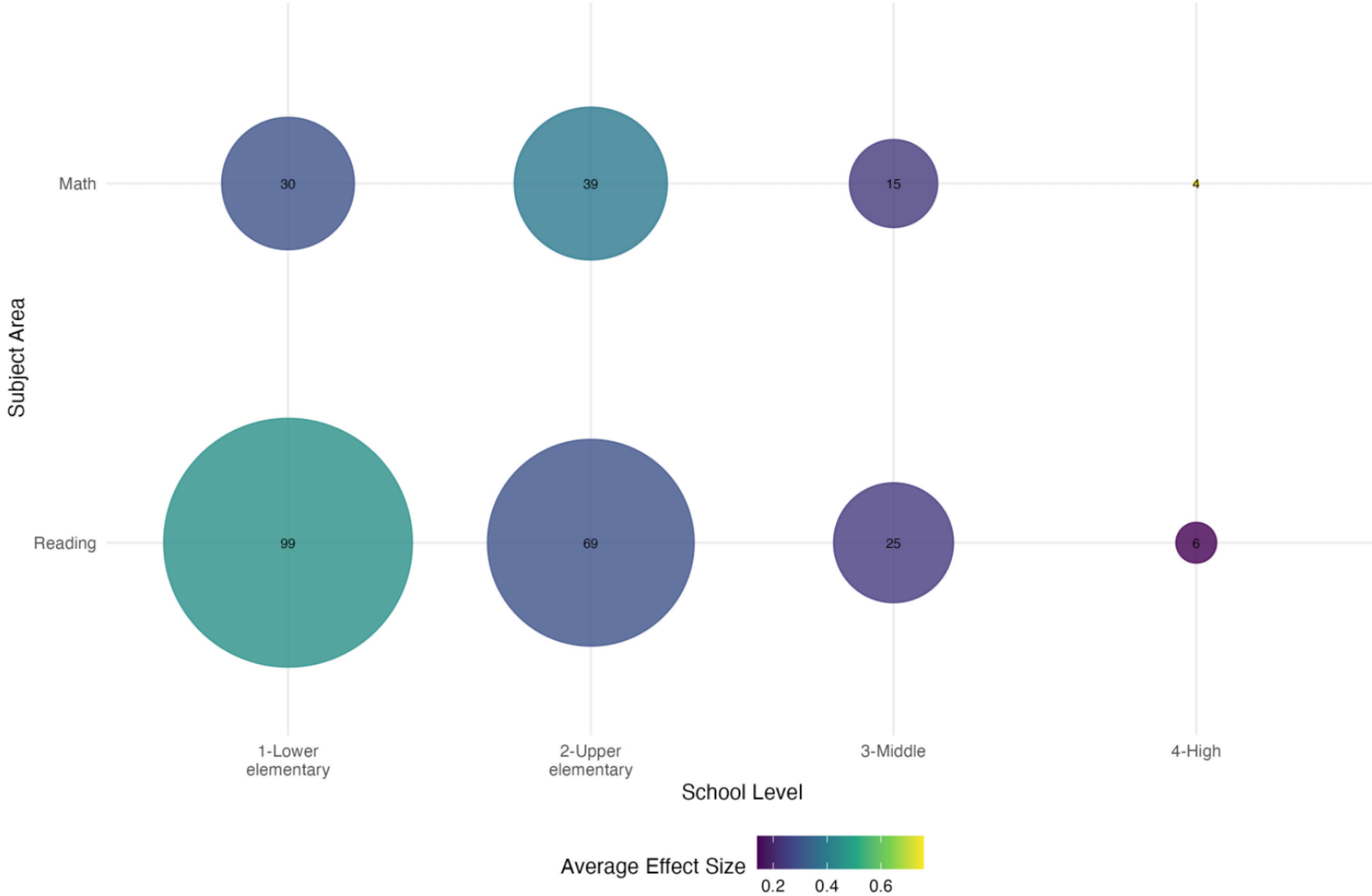
Wanzek, J., & Vaughn, S. (2008). Response to varying amounts of time in reading intervention for students with low response to intervention. *Journal of Learning Disabilities*, *41*(2), 126–142. https://doi.org/10.1177/0022219407313426

Zimmer, R., Gill, B., Razquin, P., Booker, K., & Lockwood, J. R. (2009). *State and local implementation of the No Child Left Behind Act, Volume VII--Title I school choice and Supplemental Educational Services: Final report*. U.S. Department of Education Office of Planning, Evaluation and Policy Development.

Zimmer, R., Hamilton, L., & Christina, R. (2010). After-school tutoring in the context of no Child Left Behind. *Economics of Education Review*, *29*(1), 18–28.

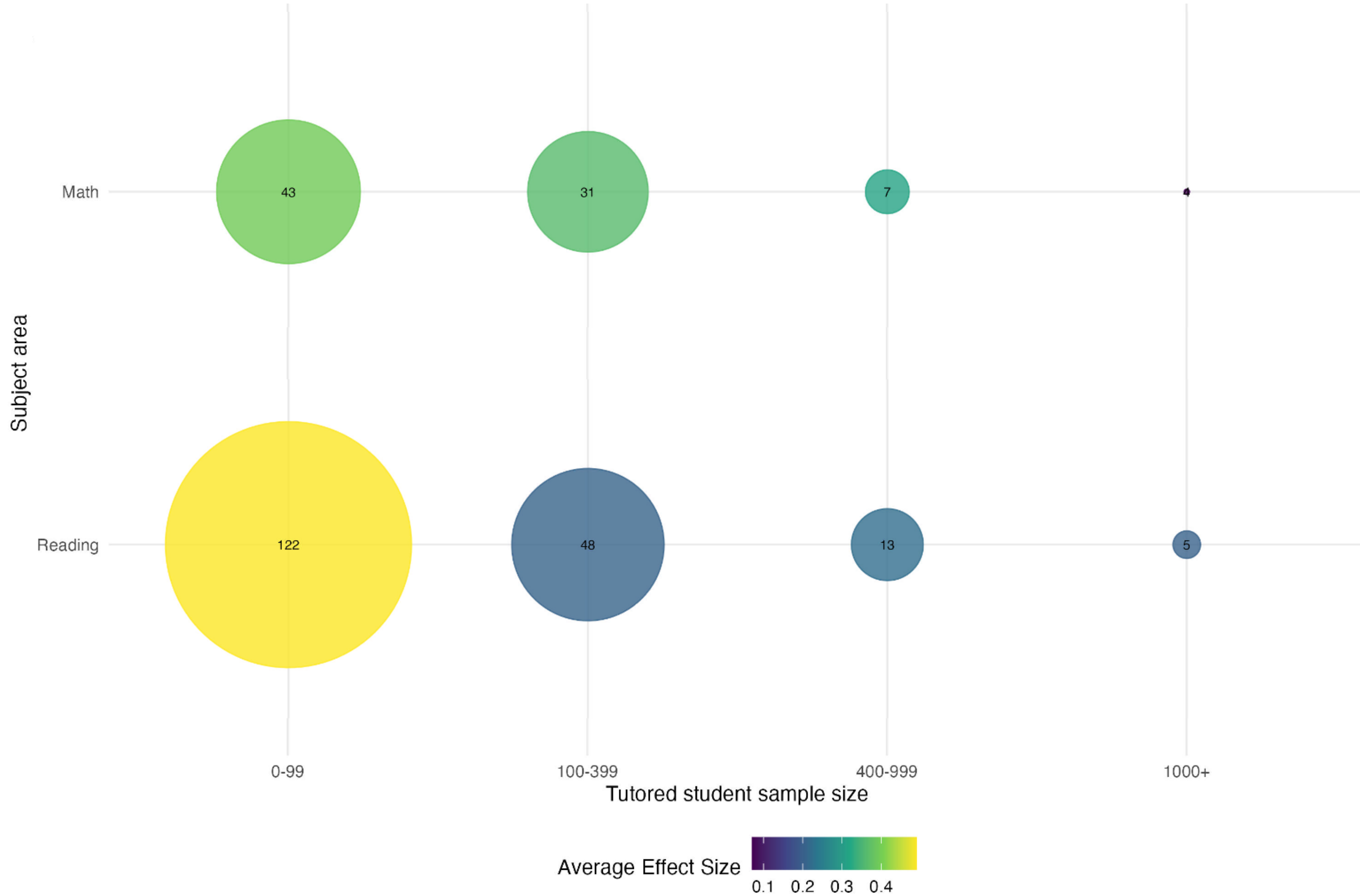
Figures and Tables

Figure 1. Evidence gap map by school level and subject area



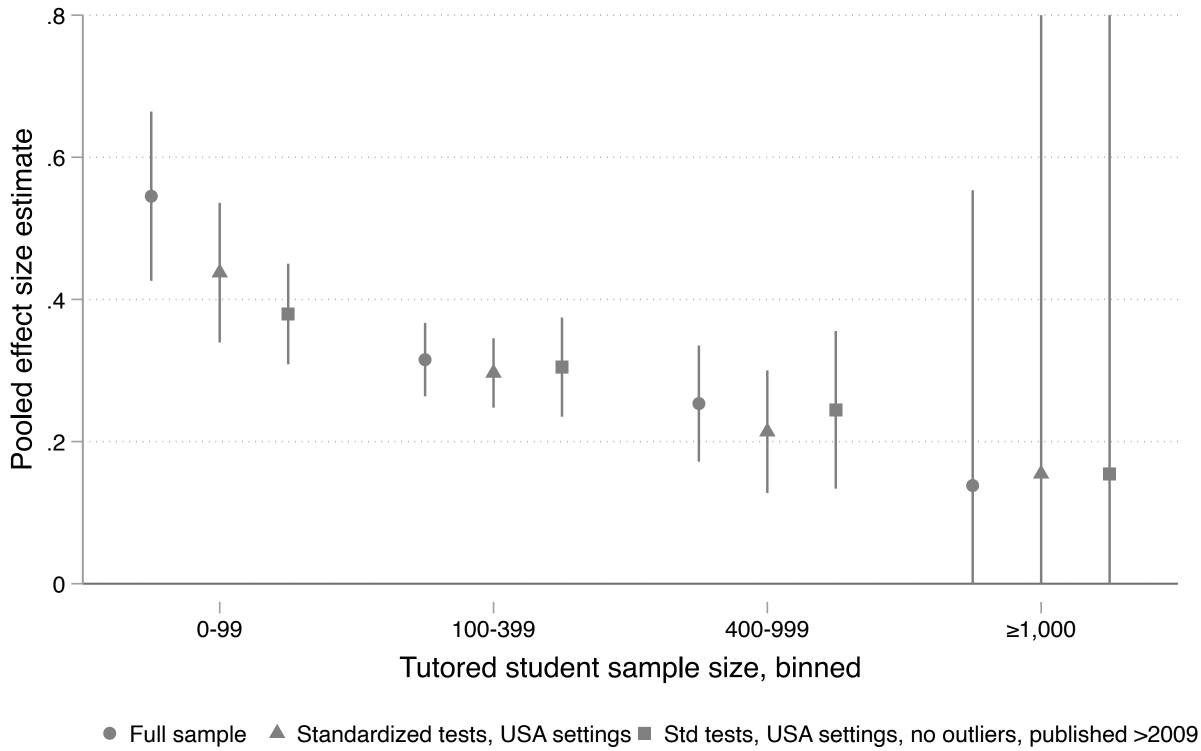
Notes: Each circle is scaled to illustrate the number of unique studies with estimates in that cross section of features. The color of each circle presents the average pooled effect size amongst those estimates. Circles are all labeled with the number of studies they represent. Note that for this figure we have established mutually exclusive categories for school level, where we round up to the higher school level observed if studies treat students are more than one level.

Figure 2. Evidence gap map by tutored student sample size and subject area



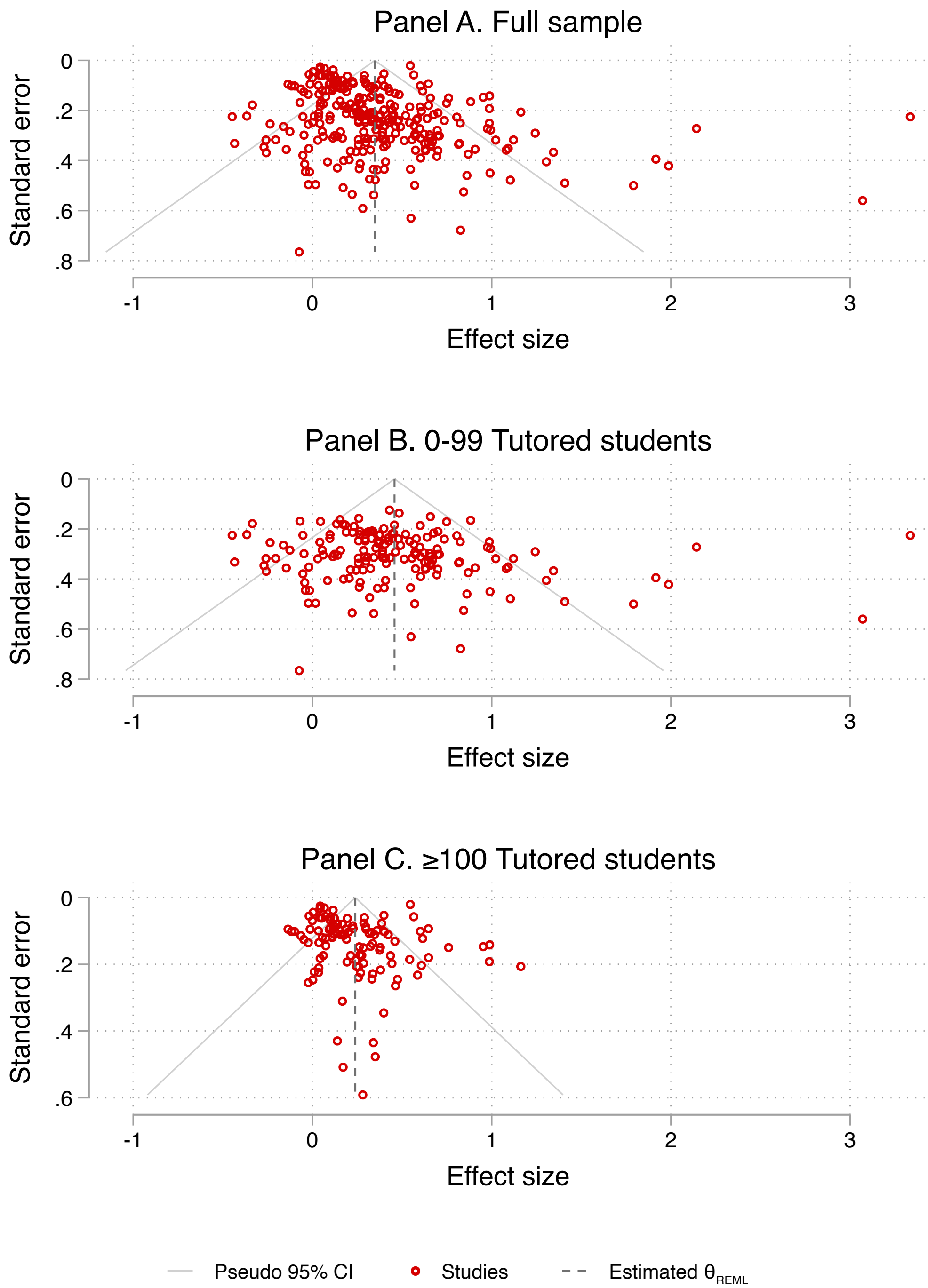
Notes: Each circle is scaled to illustrate the number of unique studies with estimates in that cross section of features. The color of each circle presents the average pooled effect size amongst those estimates. Circles are all labeled with the number of studies they represent.

Figure 3. Pooled estimated impacts of tutoring across program size and study characteristics



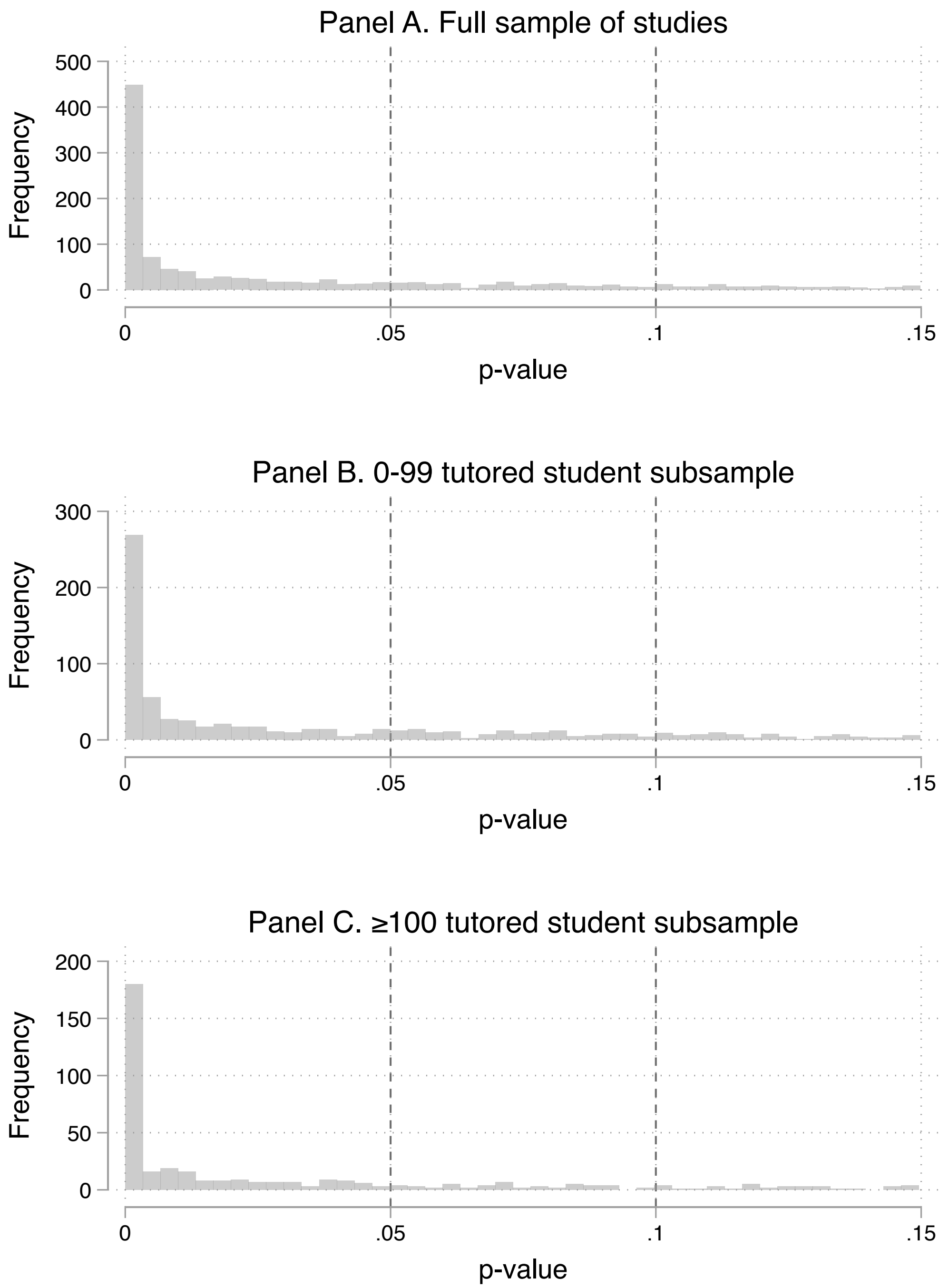
Notes. Each point represents a pooled effect size estimate for the subsample of studies triangulated by the tutored student sample size and restrictions indicated; bars represent 95% confidence intervals for these estimates. All estimates for studies with tutored student samples at or above 1,000 students are not statistically distinguishable from zero. All estimates pool impacts across math and reading.

Figure 4. Funnel plots displaying trim and fill results at the study-level



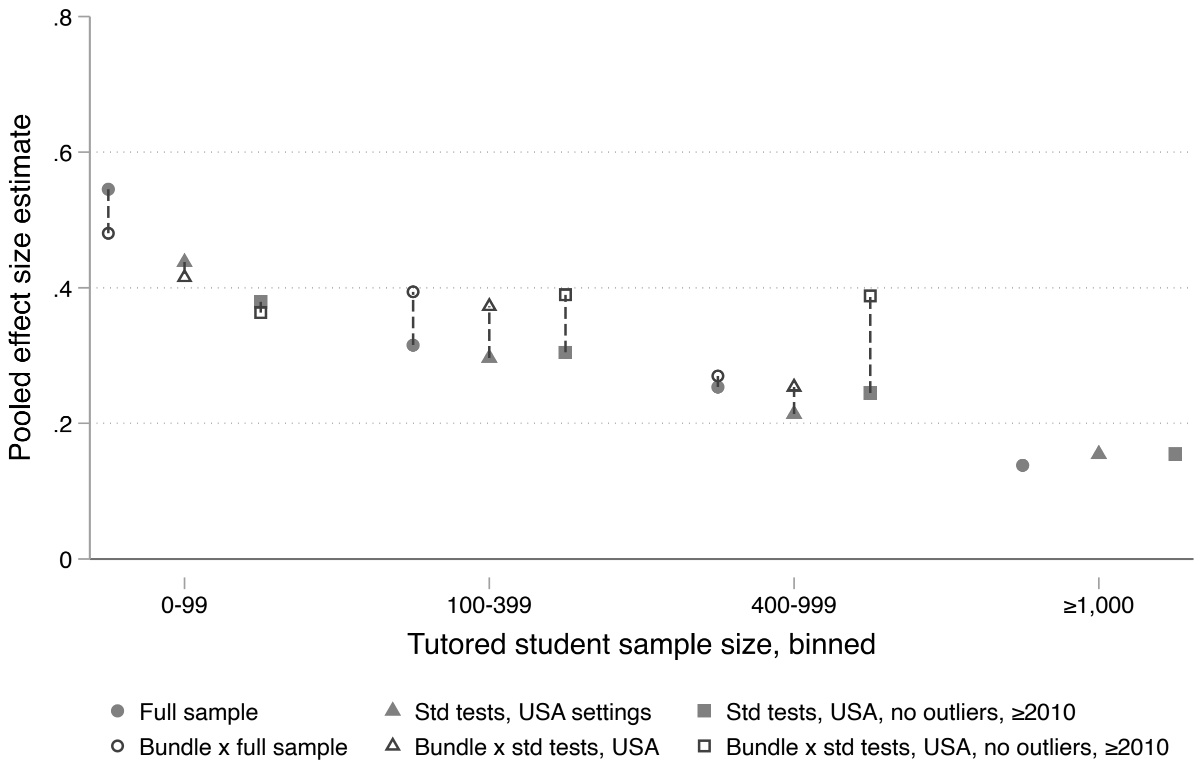
Notes: Each panel presents the average effects of each study-by-subject in the sample indicated. Panel A presents results for all studies and subjects, Panel B limits to studies of programs treating fewer than 100 students, and Panel C limits to studies of programs treating at least 100 students.

Figure 5. Density of effect estimate *p*-values across significance thresholds



Notes: Each panel presents a histogram of estimated effects' *p*-values, constrained to those <0.15, stacked across math and reading. Mass points to the left of each vertical, dashed line would suggest manipulation of results to attain significance or asymmetric reporting of results according to their significance. Panel A shows this distribution for the full sample of *p*-values below 0.15, Panels B and C disaggregate estimates according to the study's treated student sample size.

Figure 6. Gaps between pooled estimates for all studies compared to those using a bundle of tutoring best practices



Notes: Each point represents a pooled effect size estimate for the subsample of studies triangulated by the tutored student sample size and sample restrictions indicated. Dashed lines demarcate the difference between estimates using the full sample of studies indicated (solid markers) and estimates for the subsample of studies in this same group which leverage a bundle of best practices identified in the tutoring literature (hollow markers). Tutoring programs in the bundle subgroups share the following characteristics: in-person instruction, instruction at school, instruction during school, no greater than three-to-one student-to-tutor ratio, curriculum provided to tutors, at least 3 sessions per week, and at least 15 hours of total instruction dosage planned. All estimates pool across math and reading.

Table 1: Study characteristics

	Sample mean (%)	n
Publication date		
Published before 1980	3.77	10
Published in the 1980s	1.89	5
Published in the 1990s	7.92	21
Published in the 2000s	24.15	64
Published in the 2010s	50.57	134
Published in the 2020s	11.7	31
Publication type		
Peer-reviewed journal article	76.98	204
Research firm report	8.3	22
University-based research center report	1.89	5
Working paper	1.51	4
Dissertation	8.3	22
Other publication type	3.02	8
***Setting grade level**		
Lower elementary (K-2)	62.64	166
Upper elementary (3-5)	41.89	111
Middle (6-8)	12.45	33
High (9-12)	3.4	9
Setting urbanicity		
Urban setting	39.62	105
Suburban setting	4.91	13
Rural setting	4.91	13
Multiple urbanicities studied	18.49	49
Urbanicity unknown	32.08	85
Setting country		
USA	80.75	214
International / OECD Country	19.25	51
Treated student sample		
0-99 treated sample	59.25	157
100-399 treated sample	29.81	79
400-999 treated sample	7.55	20
≥1000 treated sample	3.4	9
Tutoring subject		
English as a second language	1.89	5
Math	27.55	73
Reading	64.53	171
Multiple subjects	6.04	16
N studies	265	

* Setting grade level categories are not mutually exclusive

Table 2: Intervention characteristics

	Sample mean (SD)	n
Virtual / in-person delivery		
Tutoring online	3.24	11
Tutoring in-person	96.76	329
Where tutoring happens		
Tutoring at school	85.59	291
Tutoring at home	1.18	4
Tutoring in multiple locations / other	2.65	9
Tutoring location unknown	10.59	36
When tutoring happens		
Tutoring during school	76.18	259
Tutoring after school	6.47	22
Tutoring during vacation	0.29	1
Multiple time windows / other	4.41	15
Timing unknown	12.65	43
Student-tutor ratio		
1:1 student-tutor ratio	45.88	156
2:1 student-tutor ratio	16.18	55
3:1 student-tutor ratio	13.24	45
4:1 student-tutor ratio	14.12	48
≥5:1 student-tutor ratio	7.94	27
Ratio unknown	2.65	9
Tutor type		
Tutored by teacher	17.65	60
Tutored by paraprofessional	17.06	58
Tutored by peer	9.41	32
Tutored by college / graduate student	16.18	55
Other tutor type	12.06	41
Tutor type unknown	27.65	94
***Dosage (units specified)**		
Sessions per week	3.39 (1.29)	
Hours per session	0.61 (0.38)	
Hours per week	2.01 (1.54)	
Weeks per year	16.32 (9.15)	
Hours total dosage	33.51 (31.60)	
Curriculum provided		
Yes	89.12	303
No	10.59	36
Unknown	0.29	1
N interventions	360	

* Dosage metrics are not binary variables and are not mutually exclusive. Standard deviations are reported in parentheses, where applicable. All other sets of variables are percents.

Table 3. Estimates pooled by grade level and tested subject

	Lower elementary (1)	Upper elementary (2)	Middle school (3)	High school (4)	Pooled grades (5)
Math	0.333***	0.441***	0.377***	0.550*	0.385***
	(0.036)	(0.053)	(0.075)	(0.326)	(0.032)
	[0.045, 0.620]	[0.150, 0.732]	[-0.154, 0.907]	[-0.543, 0.751]	[0.018, 0.0751]
	229	268	46	13	507
Reading	0.462***	0.480***	0.362***	0.158	0.436***
	(0.057)	(0.103)	(0.111)	(0.131)	(0.046)
	[-0.392, 1.317]	[-0.426, 1.386]	[-0.245, 0.970]	[-0.313, 0.629]	[-0.319, 1.192]
	1,263	608	127	27	1,716
Pooled subjects	0.438***	0.469***	0.360***	0.269**	0.423***
	(0.047)	(0.075)	(0.076)	(0.135)	(0.036)
	[-0.367, 1.243]	[-0.435, 1.373]	[-0.188, 0.908]	[-0.115, 0.652]	[-0.314, 1.161]
	1,492	876	173	40	2,223

Notes: *** p<0.10; ** p<0.05, * p<0.01. Prediction intervals are included for each estimate in brackets; robust standard errors are reported in parentheses. Estimates may be included in more than one group if they treat students in multiple grade levels. Lower elementary indicates treatment in grades K-2; upper elementary indicates treatment in grades 3-5; middle school indicates grades 6-8; high school indicates grades 9-12. Pooled estimates combine impact estimates for both math and reading subject tests.

Table 4. Estimates by RCT quality concerns, stacked subjects

Lower elementary (1)	Upper elementary (2)	Middle school (3)	High school (4)	Pooled grades (5)
Panel A. Studies with no RCT quality concerns				
0.437***	0.465***	0.259***	0.184**	0.415***
(0.054)	(0.081)	(0.046)	(0.090)	(0.040)
1,225	807	146	36	1,875
Panel B. Studies with an RCT quality concern				
0.436***	0.505***	1.019**		0.461***
(0.085)	(0.073)	(0.464)		(0.073)
267	69	27		348
Panel C. Omitting top and bottom 2.5% of effect size observations				
0.384***	0.376***	0.293***	0.287**	0.379***
(0.027)	(0.030)	(0.044)	(0.143)	(0.021)
1,430	815	164	39	2,113
Panel D. Studies published prior to 2000				
0.473***	0.315***	0.557***	1.198	0.449***
(0.104)	(0.115)	(0.189)	(0.779)	(0.083)
191	67	22	5	259
Panel E. Studies published between 2000 and 2009				
0.573***	0.727***	0.956**		0.584***
(0.125)	(0.233)	(0.407)		(0.103)
503	252	37		654
Panel F. Studies published between 2010 and 2019				
0.383***	0.375***	0.223***	0.178	0.376***
(0.034)	(0.046)	(0.033)	(0.122)	(0.033)
681	489	96	30	1,118
Panel G. Studies published in 2020 and following				
0.233*	0.376***	0.189**	0.116***	0.272***
(0.125)	(0.091)	(0.079)	(0.004)	(0.085)
117	68	18	5	192

Notes: *** $p<0.10$; ** $p<0.05$, * $p<0.01$. Estimates may be included in more than one group if they treat students in multiple grade levels. All cells pool across both math and reading. Cells left blank contain too few or no estimates. Panel A and B split up the entire sample by whether we identified any concerns with the quality of the RCT. Panel C omits the top and bottom 2.5% of observations by effect size magnitude.

Table 5. Pooled effect size estimates overall and by treated student sample size

No sample size restriction (1)	0-99 treated students (2)	100-399 treated students (3)	400-999 treated students (4)	≥1,000 treated students (5)
Panel A. Full analytic sample with no restrictions				
0.423***	0.545***	0.315***	0.253***	0.138
(0.036)	(0.060)	(0.026)	(0.037)	(0.103)
[-0.314, 1.161]	[-0.689, 1.780]	[-0.103, 0.733]	[0.095, 0.411]	[-0.362, 0.638]
2,223	1,403	640	112	68
Panel B. Standardized tests only				
0.352***	0.459***	0.279***	0.219***	0.138
(0.028)	(0.048)	(0.024)	(0.034)	(0.106)
[-0.228, 0.932]	[-0.359, 1.277]	[-0.148, 0.706]	[0.061, 0.377]	[-0.365, 0.642]
1,810	1,086	560	97	67
Panel C. USA settings only				
0.390***	0.463***	0.340***	0.254***	0.154
(0.027)	(0.041)	(0.027)	(0.041)	(0.148)
[-0.228. 0.932]	[-0.348, 1.275]	[-0.030, 0.709]	[0.071, 0.437]	[-0.463, 0.771]
1,829	1,152	528	90	59
Panel D. Standardized tests only, USA settings only				
0.351***	0.438***	0.297***	0.214***	0.155
(0.029)	(0.049)	(0.024)	(0.037)	(0.154)
[-0.245, 0.947]	[-0.347, 1.222]	[-0.066. 0.660]	[0.055, 0.373]	[-0.468, 0.778]
1,484	895	456	75	58
Panel E. Omitting the top and bottom 2.5% of effect sizes				
0.379***	0.464***	0.307***	0.253***	0.138
(0.021)	(0.028)	(0.025)	(0.037)	(0.103)
[-0.193, 0.951]	[-0.044, 0.973]	[-0.078, 0.692]	[0.095, 0.411]	[-0.362, 0.638]
2,113	1,306	627	112	68
Panel F. Studies published in or since 2010				
0.357***	0.433***	0.328***	0.289***	0.138
(0.031)	(0.053)	(0.035)	(0.039)	(0.103)
[-0.263, 0.976]	[-0.337, 1.204]	[-0.169, 0.825]	[0.202, 0.375]	[-0.362, 0.638]
1,310	725	430	87	68
Panel G. Standardized tests only, USA settings only, omitting the top and bottom 2.5% of effect sizes, studies published in or since 2010				
0.320***	0.379***	0.305***	0.245***	0.155
(0.030)	(0.034)	(0.034)	(0.045)	(0.154)
[-0.054, 0.694]	[0.058, 0.701]	[-0.054, 0.663]	[0.095, 0.394]	[-0.468, 0.778]
840	462	268	52	58

Notes: *** p<0.10; ** p<0.05, * p<0.01. Prediction intervals are included for each estimate in brackets; robust standard errors are reported in parentheses. Each cell presents the Hedges' g estimate stacking both math and reading. Model (1) offers the pooled average impact of tutoring across the entire subsample indicated in each panel. Models (2) through (5) disaggregate the estimate in Model (1) by the tutored student sample size of each study. Panel D combines the sample restrictions in Panels B and C; Panel G combines the sample restrictions in Panels D, E, and F.

Table 6. Tests for significant differences in estimate mass across p-value thresholds

	p-value threshold = 0.10			*p*-value threshold = 0.05		
Bandwidth	± 0.02	± 0.01	± 0.005	± 0.02	± 0.01	± 0.005
Panel A. Full sample						
N estimates within bandwidth	117	54	32	180	91	47
% significant estimates	0.52	0.50	0.41	0.56	0.48	0.45
One-sided *p*-value	0.36	0.55	0.89	0.06	0.66	0.81
Panel B. Studies with 0-99 treated students						
N estimates within bandwidth	85	42	25	123	64	35
% significant estimates	0.52	0.50	0.44	0.53	0.42	0.43
One-sided *p*-value	0.41	0.56	0.79	0.29	0.92	0.84
Panel C. Studies with 100+ treated students						
N estimates within bandwidth	32	12	7	57	27	12
% significant estimates	0.53	0.50	0.29	0.63	0.63	0.50
One-sided *p*-value	0.43	0.61	0.94	0.03	0.12	0.61

Notes: Here we present the likelihood of observing the number of significant *p*-values in our data at the 5% and 10% significance levels within the bandwidths 0.02, 0.01, and 0.005 around those thresholds. For each of these combinations, we isolate the subsample within the indicated bandwidth around the indicated threshold ("N estimates within bandwidth"), present the share of significant estimate *p*-values in that range ("Share significant estimates"), calculate the likelihood of having at least that many significant estimates assuming a binomial distribution ("One-sided *p*-value"). We repeat this exercise for our full sample of estimates in Panel A, and disaggregated according to treated student sample size in Panels B and C. All estimates pool across both math and reading subject areas. All cells pool across math and reading.

Table 7. Intervention characteristics across treated student sample size

	Treated student sample size			
	0 to 99	100 to 399	400 to 999	≥1,000
Virtual / in-person delivery				
Tutoring online	0.52	6.31	0.00	23.08
Tutoring in-person	99.48	93.69	100.00	76.92
Where tutoring happens				
Tutoring at school	86.01	81.08	95.65	100.00
Tutoring at home	0.52	2.70	0.00	0.00
Tutoring in multiple locations / other	2.59	3.60	0.00	0.00
Tutoring location unknown	10.88	12.61	4.35	0.00
When tutoring happens				
Tutoring during school	76.68	72.07	86.96	84.62
Tutoring after school	6.22	9.01	0.00	0.00
Tutoring during vacation	0.52	0.00	0.00	0.00
Multiple time windows / other	3.11	6.31	8.70	0.00
Timing unknown	13.47	12.61	4.35	15.38
Student-tutor ratio				
1:1 student-tutor ratio	49.22	42.34	39.13	38.46
2:1 student-tutor ratio	11.92	21.62	13.04	38.46
3:1 student-tutor ratio	12.44	15.32	17.39	0.00
4:1 student-tutor ratio	13.99	16.22	13.04	0.00
≥5:1 student-tutor ratio	9.84	2.70	8.70	23.08
Ratio unknown	2.59	1.80	8.70	0.00
Tutor type				
Tutored by teacher	15.03	19.82	21.74	30.77
Tutored by paraprofessional	15.03	18.02	26.09	23.08
Tutored by peer	13.99	1.80	4.35	15.38
Tutored by college / grad student	19.17	14.41	4.35	7.69
Other tutor type	7.25	16.22	26.09	23.08
Tutor type unknown	29.53	29.73	17.39	0.00
Dosage				
Sessions per week	3.31	3.45	3.55	3.58
	(1.35)	(1.09)	(1.58)	(1.68)
Hours per session	0.61	0.60	0.53	0.72
	(0.44)	(0.29)	(0.23)	(0.39)
Hours per week	1.98	2.03	1.90	2.30
	(1.82)	(1.06)	(1.15)	(1.33)
Weeks per year	15.02	18.61	18.48	13.71
	(9.74)	(7.90)	(8.26)	(3.30)
Hours total dosage	30.32	39.17	36.85	27.00
	(31.68)	(31.00)	(35.10)	(11.84)
N interventions	193	111	23	13

Notes: Except for dosage variables, all measures are percents scaled 0 to 100. Dosage variable units are indicated, with standard deviations presented in parentheses.

Table 8. Meta-regression controlling for study and intervention features

	(1)		(2)	
100-399 treated sample (ref. 0-99)	-0.203***	(0.062)	-0.140**	(0.058)
400-999 treated sample (ref. 0-99)	-0.272***	(0.067)	-0.164**	(0.069)
≥1,000 treated sample (ref. 0-99)	-0.386***	(0.117)	-0.284*	(0.161)
Published in 2000s (ref. pre-2000)			0.152	(0.157)
Published in 2010s (ref. pre-2000)			0.004	(0.131)
Published in 2020s (ref. pre-2000)			0.054	(0.151)
Flag for poor RCT quality			0.005	(0.076)
Lower elementary math (ref. LE reading)			-0.150*	(0.084)
Upper elementary reading (ref. LE reading)			-0.029	(0.065)
Upper elementary math (ref. LE reading)			-0.122	(0.104)
Middle school reading (ref. LE reading)			-0.073	(0.140)
Middle school math (ref. LE reading)			-0.096	(0.138)
High school reading (ref. LE reading)			-0.224	(0.150)
High school math (ref. LE reading)			0.295	(0.267)
Researcher-generated assessment (ref. standardized test)			0.164*	0.224**
OECD country (ref. USA)			0.135	(0.142)
Tutoring delivered online (ref. in-person)			0.034	(0.171)
Tutoring at multiple locations / other (ref. at school)			-0.148	-0.016
Tutoring location missing (ref. at school)			0.069	(0.189)
Curriculum not provided			0.135	(0.148)
Tutoring outside of school hours (ref. during school)			-0.165***	-0.192***
Tutoring timing missing (ref. during school)			0.16	(0.151)
2:1 student-tutor ratio (ref. 1:1)			0.053	(0.063)
3:1 student-tutor ratio (ref. 1:1)			-0.058	(0.087)
4:1 student-tutor ratio (ref. 1:1)			-0.059	(0.077)
≥5:1 student-tutor ratio (ref. 1:1)			0.237	(0.251)
Ratio missing (ref. 1:1)			-0.186*	(0.102)
Tutored by paraprofessional (ref. teacher)			-0.042	(0.088)
Tutored by K-12 peer (ref. teacher)			-0.056	(0.125)
Tutored by college / graduate student (ref. teacher)			0.071	-0.081
Other tutor type (ref. teacher)			-0.162*	(0.088)
Tutor type missing (ref. teacher)			0.084	(0.091)
Total dosage 0-14 hours (ref. ≥60 hours)			0.17	(0.117)
Total dosage 15-29 hours (ref. ≥60 hours)			0.206**	(0.092)
Total dosage 30-44 hours (ref. ≥60 hours)			0.15	(0.106)
Total dosage 45-59 hours (ref. ≥60 hours)			0.213**	(0.083)
Total dosage missing (ref. ≥60 hours)			0.057	(0.089)
Constant	0.523***	(0.056)	0.274	(0.167)
Observations	2,223		2,223	

Notes: *** p<0.10; ** p<0.05, * p<0.01. Standard errors are presented in parentheses.

Table 9. Student sample characteristics by size of tutored student sample

	Treated student sample size	
	0 to 99	≥100
Student Demographics		
% Asian	2.48	2.40
% Black	31.04	33.52
% Hispanic / Latinx	29.41	24.98
% Native American	2.41	0.52
% Multiracial	0.59	2.26
% White	28.71	35.59
% Other	6.23	7.29
% free/reduced-price lunch	72.83	65.17
% special education	28.80	17.75
% English language learners	30.70	17.15
Program Targeted to Certain Students		
Any targeting	92.75	95.92
Targets low performers	88.60	85.71
Targets ELLs	8.29	0.68
Targets underserved students	19.17	21.77
Targets socioemotional problems	3.11	3.40
N interventions	193	147

Notes: Means are taken at the intervention level. All variables presented range from 0 to 100.

Table 10. Estimates for programs that combine best practices, stacked subjects

No sample size restriction	0-99 treated students	100-399 treated students	400-999 treated students
(1)	(2)	(3)	(4)
Panel A. Subsample of programs in-person, at school, during school, with ratio no more than 3:1, provided curricula, meeting ≥3 times per week, ≥15 hours of dosage			
0.433***	0.480***	0.394***	0.270***
(0.039)	(0.059)	(0.046)	(0.068)
732	474	216	42
Panel B. Subsample of programs using standardized tests, USA settings, in-person, at school, during school, with ratio no more than 3:1, provided curricula, meeting ≥3 times per week, ≥15 hours of dosage			
0.385***	0.415***	0.373***	0.254***
(0.036)	(0.053)	(0.049)	(0.073)
519	336	147	36
Panel C. Subsample of programs standardized tests, USA settings, no outliers, published since 2010, in-person, at school, during school, with ratio no more than 3:1, provided curricula, meeting ≥3 times per week, ≥15 hours of dosage			
0.375***	0.363***	0.390***	0.388***
(0.037)	(0.059)	(0.056)	(0.066)
299	180	103	16

Notes: *** p<0.10; ** p<0.05, * p<0.01. All cells stack estimates for math and reading. Column (1) presents the pooled meta-analytic estimated effect for the subsample of studies described in each panel. Columns (2) through (4) disaggregate the estimate in Column (1) according to the tutored student sample size. Panel A limits to the described subsample of programs sharing a set of best practices in their designs. Panel B restricts to studies with standardized test outcome measures and conducted in USA settings. Panel C additionally drops the top and bottom 2.5% of effect sizes from the whole sample ("no outliers"), and excludes studies published prior to 2010.

Table 11a. Average effect sizes across different program design features, stacked subjects

Delivery mode		Student-tutor ratio				
In-person	Virtual	1:1 ratio	2:1 ratio	3:1 ratio	4:1 ratio	≥5:1 ratio
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Full analytic sample with no restrictions						
0.438***	0.065**	0.432***	0.406***	0.299***	0.341***	0.909**
(0.036)	(0.028)	(0.040)	(0.052)	(0.045)	(0.068)	(0.364)
2,164	59	1,061	290	314	367	154
Panel B. Standardized tests only, USA settings only						
0.368***	0.052***	0.465***	0.298***	0.263***	0.248***	0.244***
(0.028)	(0.005)	(0.054)	(0.056)	(0.040)	(0.041)	(0.083)
1,438	46	685	189	246	282	62
Panel C. Standardized tests only, USA settings only, omitting the top and bottom 2.5% of effect sizes, studies published in or since 2010						
0.345***	0.052***	0.374***	0.268***	0.292***	0.286***	0.292***
(0.023)	(0.005)	(0.053)	(0.094)	(0.039)	(0.042)	(0.076)
794	46	319	87	204	173	42

Notes: *** $p<0.10$; ** $p<0.05$, * $p<0.01$. Each column isolates a subsample of effects according to tutoring program characteristics. All estimates stack math and reading.

Table 11b. Average effect sizes across different program design features, stacked subjects

Tutor type				Total dosage of tutoring hours				
Teacher	Paraeducator	College / graduate student	K-12 peer	0-14 hours	15-29 hours	30-44 hours	45-59 hours	≥60 hours
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Full analytic sample with no restrictions								
0.381***	0.428***	0.397***	0.358***	0.541***	0.484***	0.465***	0.422***	0.256***
(0.055)	(0.042)	(0.049)	(0.092)	(0.133)	(0.069)	(0.067)	(0.074)	(0.044)
496	363	328	122	481	520	399	158	342
Panel B. Standardized tests only, USA settings only								
0.309***	0.396***	0.358***	0.429***	0.313***	0.507***	0.348***	0.430***	0.241***
(0.044)	(0.046)	(0.060)	(0.127)	(0.042)	(0.102)	(0.050)	(0.091)	(0.049)
388	229	194	93	281	323	236	109	291
Panel C. Standardized tests only, USA settings only, omitting the top and bottom 2.5% of effect sizes, studies published in or since 2010								
0.338***	0.355***	0.375***	0.109	0.308***	0.412***	0.305***	0.398***	0.220***
(0.046)	(0.040)	(0.075)	(0.189)	(0.045)	(0.042)	(0.044)	(0.109)	(0.066)
200	140	133	19	208	207	116	62	124

Notes: *** p<0.10; ** p<0.05, * p<0.01. Each column isolates a subsample of effects according to tutoring program characteristics. All estimates stack math and reading.

Table 12a. Multi-arm studies experimentally comparing different student-tutor ratios

Citation	N Treated Students	Subject	Small Ratio	Large Ratio	Students per tutor													Diff. (Small - Big)
					1	2	3	4	5	6	7	8	9	10	11	12	13	
Carlana & La Ferrara, 2024	607	Multiple	1:1	2:1														0.09
Clarke et al., 2017	415	Math	2:1	5:1														0.09
																		0.52
																		0.14
																		0.25
																		0.76
Clarke et al., 2020	880	Math	2:1	5:1														-0.02
																		-0.03
																		-0.03
																		0.01
																		-0.03
																		0.12
Clarke et al., 2023	322	Math	2:1	5:1														0.07
																		0.20
																		0.77
Doabler et al., 2019	465	Math	2:1	5:1														-0.01
																		0.04
																		0.00
																		0.10
																		-0.01
Kraft & Lovison, 2024	180	Math	1:1	3:1														0.14
Loeb et al., 2023	1,080	Reading	1:1	2:1														0.06
																		0.03

Table 12a. Continued

Citation	N Treated Students	Subject	Small Ratio	Large Ratio	Students per tutor 1	2	3	4	5	6	7	8	9	10	11	12	13	Diff. (Small - Big)
Schwartz et al., 2012	170	Reading	1:1	3:1														0.63
																		0.35
																		0.41
																		0.23
																		0.41
																		0.30
																		0.36
Vadasy & Sanders, 2008	54	Reading	1:1	2:1														-0.09
																		-0.22
																		-0.37
																		-0.12
																		-0.06
																		-0.22
																		-0.08
Vaughn et al., 2010	514	Reading	4:1	13:1														-0.10
																		-0.01
																		0.07
																		0.11
																		-0.08
																		0.26
																		0.08
																		0.22
																		0.13
																		0.11
																		0.06
																		0.23

Notes: All studies examine elementary programs except for three that study middle school programs: Carlana & La Ferrara (2024), Kraft & Lovison (2024) & Vaughn et al. (2010)

Table 12b. Multi-arm studies experimentally comparing different dosages of tutoring

Citation	N of Treated Students	Amount of Low Dosage	Amount of High Dosage	Total hours of tutoring												Effect Size Diff. (High - Low)
				2	8	14	20	26	32	38	44	50	56	62	68	
Al Otaiba et al., 2005	49	Two 20 min. sessions / week	Four 20 min. sessions / week													0.01 0.01 -0.12 0.18 0.13 0.15 0.18
Begeny, 2011	58	1.5 nine min. sessions / week	Three nine min. sessions / week													0.10 0.29
Carlana & La Ferrara, 2021	530	Three hours / week	Six hours / week													0.22
Wanzek & Vaughn, 2008	35	Five 30 min. sessions / week	Ten 30 min. sessions / week													-0.4 0.54 0.39 -0.76

Note: There is more than one effect in each study because authors report effects on multiple reading outcomes or assessments. All studies examine tutoring in reading subjects except for Carlana & La Ferrera (2021) which focused on multiple subjects.

Appendix A. Included Studies

Al Otaiba, S., Schatschneider, C., & Silverman, E. (2005). Tutor-assisted intensive learning strategies in kindergarten: How much is enough? *Exceptionality*, *13*(4), 195–208. https://doi.org/10.1207/s15327035ex1304_2

Allen, A. A., & Lembke, E. S. (2022). The effect of a morphological awareness intervention on early writing outcomes. *Learning Disability Quarterly*, *45*(2), 72–84. https://doi.org/10.1177/0731948720912414

Allor, J. H., Mathes, P. G., Roberts, J. K., Jones, F. G., & Champlin, T. M. (2010). Teaching students with moderate intellectual disabilities to read: An experimental examination of a comprehensive reading intervention. *Education and Training in Autism and Developmental Disabilities*, *45*(1), 3–22.

Allor, J., & McCathren, R. (2004). The efficacy of an early literacy tutoring program implemented by college students. *Learning Disabilities Research & Practice*, *19*(2), 116–129. https://doi.org/10.1111/j.1540-5826.2004.00095.x

Altman, J. R., D'Brot, J. M., & Akhtar, N. M. (2016). *America Reads - Mississippi: An impact evaluation of the program's 2015-2016 school year*. TerraLuna Collective.

Amendum, S. J., Vernon-Feagans, L., & Ginsberg, M. C. (2011). The effectiveness of a technologically facilitated classroom-based early reading intervention. *The Elementary School Journal*, *112*(1), 107–131. https://doi.org/10.1086/660684

Anderberg, R. G. (2013). *The effects of cross-age tutoring on the reading ability of first and second grade students*. Middle Tennessee State University.

Apel, K., & Diehm, E. (2014). Morphological awareness intervention with kindergarteners and first and second grade students from low ses homes: A small efficacy study. *Journal of Learning Disabilities*, *47*(1), 65–75. https://doi.org/10.1177/0022219413509964

Baker, D. L., Burns, D., Kame'enui, E. J., Smolkowski, K., & Baker, S. K. (2016). Does supplemental instruction support the transition from Spanish to English reading instruction for first-grade english learners at risk of reading difficulties? *Learning Disability Quarterly*, *39*(4), 226–239. https://doi.org/10.1177/0731948715616757

Baker, S., Gersten, R., & Keating, T. (2000). When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training. *Reading Research Quarterly*, *35*(4), 494–519. https://doi.org/10.1598/RRQ.35.4.3

Ball, E. W., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly*, *26*(1), 49–66.

Bausell, R. B., Moody, W. B., & Walzl, F. N. (1972). A factorial study of tutoring versus classroom instruction. *American Educational Research Journal*, *9*(4), 591–597. https://doi.org/10.3102/00028312009004591

Begeny, J. C. (2011). Effects of the Helping Early Literacy with Practice Strategies (helps) reading fluency program when implemented at different frequencies. *School Psychology Review*, *40*(1), 149–157. https://doi.org/10.1080/02796015.2011.12087734

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology*, *96*(3), 444–461. https://doi.org/10.1037/0022-0663.96.3.444

Blackburn, S. (2014). *The effect of the HELPS program on the oral reading fluency and accuracy rates of third-, fourth-, and fifth-grade students* [Gardner-Webb University]. https://digitalcommons.gardner-webb.edu/education_etd/4

Bøg, M., Dietrichson, J., & Aldenius, A. (2019). *A multi-sensory tutoring program for students at-risk of reading difficulties: Evidence from a randomized field experiment* (Working Paper 2019:7). Working Paper. https://www.econstor.eu/handle/10419/201471

Bonesrønning, H., Finseraas, H., Hardoy, I., Iversen, J. M. V., Nyhus, O. H., Opheim, V., Salvanes, K. V., Sandsør, A. M. J., & Schøne, P. (2022). Small-group instruction to improve student performance in mathematics in early grades: Results from a randomized field experiment. *Journal of Public Economics*, *216*, 104765. https://doi.org/10.1016/j.jpubeco.2022.104765

Borman, G. D., Borman, T. H., Park, S. J., & Houghton, S. (2020). A multisite randomized controlled trial of the effectiveness of Descubriendo la Lectura. *American Educational Research Journal*, *57*(5), 1995–2020. https://doi.org/10.3102/0002831219890612

Bramuchi, L. H. (2009). *Reading interventions to improve: Oral reading fluency and literal comprehension* [Ed.D.]. Delta State University.

Bryant, D. P., Bryant, B. R., Roberts, G., Vaughn, S., Pfannenstiel, K. H., Porterfield, J., & Gersten, R. (2011). Early numeracy intervention program for first-grade students with mathematics difficulties. *Exceptional Children*, *78*(1), 7–23. https://doi.org/10.1177/001440291107800101

Buckingham, J., Wheldall, K., & Beaman, R. (2012). A randomised control trial of a Tier-2 small-group intervention (‘MiniLit’) for young struggling readers1. *Australian Journal of Learning Difficulties*, *17*(2), 79–99. https://doi.org/10.1080/19404158.2012.717537

Cabezas, V., Cuesta, J. I., & Gallego, F. A. (2011). *Effects of short-term tutoring on cognitive and non-cognitive skills: Evidence from a randomized evaluation in Chile.* Abdul Latif Jameel Poverty Action Lab. https://www.povertyactionlab.org/sites/default/files/research-paper/493%20-%20short-term%20tutoring%20May2011.pdf

Calhoon, M. B., Al Otaiba, S., Greenberg, D., King, A., & Avalos, A. (2006). Improving reading skills in predominantly Hispanic title 1 first-grade classrooms: The promise of Peer-Assisted Learning Strategies. *Learning Disabilities Research & Practice*, *21*(4), 261–272. https://doi.org/10.1111/j.1540-5826.2006.00222.x

Carlton, M. B., Litton, F. W., & Zinkgraf, S. A. (1985). The effects of an intraclass peer tutoring program on the sight-word recognition ability of students who are mildly mentally retarded. *Mental Retardation*, *23*(2), 74–78.

Case, L., Speece, D., Silverman, R., Schatschneider, C., Montanaro, E., & Ritchey, K. (2014). Immediate and long-term effects of tier 2 reading instruction for first-grade students with a high probability of reading failure. *Journal of Research on Educational Effectiveness*, *7*(1), 28–53. https://doi.org/10.1080/19345747.2013.786771

Center, Y., Wheldall, K., Freeman, L., Outhred, L., & McNaught, M. (1995). An evaluation of Reading Recovery. *Reading Research Quarterly*, *30*(2), 240–263. https://doi.org/10.2307/748034

Cirino, P. T., Vaughn, S., Linan-Thompson, S., Cardenas-Hagan, E., Fletcher, J. M., & Francis, D. J. (2009). One-year follow-up outcomes of Spanish and English interventions for English language learners at risk for reading problems. *American Educational Research Journal*, *46*(3), 744–781. https://doi.org/10.3102/0002831208330214

Clarke, B., Doabler, C., Smolkowski, K., Kurtz Nelson, E., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the immediate and long-term efficacy of a tier 2 kindergarten mathematics intervention. *Journal of Research on Educational Effectiveness*, *9*(4), 607–634. https://doi.org/10.1080/19345747.2015.1116034

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), 233285841770689. https://doi.org/10.1177/2332858417706899

Clarke, B., Doabler, C. T., Smolkowski, K., Baker, S. K., Fien, H., & Strand Cary, M. (2016). Examining the efficacy of a tier 2 kindergarten mathematics intervention. *Journal of Learning Disabilities*, *49*(2), 152–165. https://doi.org/10.1177/0022219414538514

Clarke, B., Doabler, C. T., Turtura, J., Smolkowski, K., Kosty, D. B., Sutherland, M., Kurtz-Nelson, E., Fien, H., & Baker, S. K. (2020). Examining the efficacy of a kindergarten mathematics intervention by group size and initial skill: Implications for practice and policy. *The Elementary School Journal*, *121*(1), 125–153. https://doi.org/10.1086/710041

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*(8), 1106–1116. https://doi.org/10.1177/0956797610375449

Cloward, R. D. (1967). Studies in tutoring. *The Journal of Experimental Education*. https://www.tandfonline.com/doi/abs/10.1080/00220973.1967.11011022

Codding, R. S., Nelson, P. M., Parker, D. C., Edmunds, R., & Klaft, J. (2022). Examining the impact of a tutoring program implemented with community support on math proficiency and growth. *Journal of School Psychology*, *90*, 82–93. https://doi.org/10.1016/j.jsp.2021.11.002

Cook, J. A. (2001). *Every moment counts: Pairing struggling young readers with minimally trained tutors.* Arizona State University.

Cook, P. J., Dodge, K., Farkas, G., Fryer, J., Roland G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H., & Steinberg, L. (2014). *The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in chicago* (Working Paper No. 19862). National Bureau of Economic Research. https://doi.org/10.3386/w19862

Cortes, K., Kortecamp, K., Loeb, S., & Robinson, C. (2023). *A scalable approach to high-impact tutoring for young readers: Results of a randomized controlled trial* [Working Paper]. National Student Support Accelerator.

Courtney, M. E., Zinn, A., Zielewski, E. H., Bess, R. J., Malm, K. E., Stagner, M., & Pergamit, M. (2008). Evaluation of the Early Start to Emancipation Preparation Tutoring Program, Los Angeles County, California: Final report. In *Administration for Children & Families*. Administration for Children & Families. https://eric.ed.gov/?id=ED502639

Coyne, M. D., McCoach, D. B., Ware, S., Austin, C. R., Loftus-Rattan, S. M., & Baker, D. L. (2019). Racing against the vocabulary gap: Matthew effects in early vocabulary instruction and intervention. *Exceptional Children*, *85*(2), 163–179. https://doi.org/10.1177/0014402918789162

Dahlem, G. G. (1973). *The effect of like ethnic qualities upon reading tutoring of third graders. final report*. https://eric.ed.gov/?id=ED095488

De Ree, J., Maggioni, M. A., Paulle, B., Rossignoli, D., Ruijs, N., & Walentek, D. (2023). Closing the income-achievement gap? Experimental evidence from high-dosage tutoring in Dutch primary education. *Economics of Education Review*, *94*, 102383. https://doi.org/10.1016/j.econedurev.2023.102383

De Ree, J., Maggioni, M. A., Paulle, B., Rossignoli, D., & Walentek, D. (2021). *High dosage tutoring in pre-vocational secondary education: Experimental evidence from Amsterdam*. OSF. https://doi.org/10.31235/osf.io/r56um

Denton, C. A., Anthony, J. L., Parker, R., & Hasbrouck, J. E. (2004). Effects of two tutoring programs on the English reading development of Spanish-English bilingual students. *The Elementary School Journal*, *104*(4), 289–305. https://doi.org/10.1086/499754

Denton, C. A., Fletcher, J. M., Taylor, W. P., Barth, A. E., & Vaughn, S. (2014). An experimental evaluation of guided reading and explicit interventions for primary-grade students at-risk for reading difficulties. *Journal of Research on Educational Effectiveness*, *7*(3), 268–293. https://doi.org/10.1080/19345747.2014.906010

Denton, C. A., Nimon, K., Mathes, P. G., Swanson, E. A., Kethley, C., Kurz, T. B., & Shih, M. (2010). Effectiveness of a supplemental early reading intervention scaled up in multiple schools. *Exceptional Children*, *76*(4), 394–416. https://doi.org/10.1177/001440291007600402

Denton, C. A., Tolar, T. D., Fletcher, J. M., Barth, A. E., Vaughn, S., & Francis, D. J. (2013). Effects of tier 3 intervention for students with persistent reading difficulties and characteristics of inadequate responders. *Journal of Educational Psychology*, *105*(3), 633–648. https://doi.org/10.1037/a0032581

Denton, C. A., Wexler, J., Vaughn, S., & Bryan, D. (2008). Intervention provided to linguistically diverse middle school students with severe reading difficulties. *Learning Disabilities Research & Practice*, *23*(2), 79–89. https://doi.org/10.1111/j.1540-5826.2008.00266.x

Doabler, C. T., Clarke, B., Kosty, D. B., Kurtz-Nelson, E., Fien, H., Smolkowski, K., & Baker, S. K. (2016). Testing the efficacy of a tier 2 mathematics intervention: A conceptual replication study. *Exceptional Children*, *83*(1), 92–110. https://doi.org/10.1177/0014402916660084

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities*, *46*(2), 166–181. https://doi.org/10.1177/0022219411410233

Eddy, R. M., Ruitman, H. T., Hankel, N., Matelski, M. H., & Schmalstig, M. (2011). *Pearson words their way: Word study in action*. Cobblestone Applied Research & Evaluation, Inc.

Ehri, L. C., Dreyer, L. G., Flugman, B., & Gross, A. (2007). Reading Rescue: An effective tutoring intervention model for language-minority students who are struggling readers in first grade. *American Educational Research Journal*, *44*(2), 414–448. https://doi.org/10.3102/0002831207302175

Esser, M. M. S. (2001). *The effects of metacognitive strategy training and attribution retraining on reading comprehension in African -American students with learning disabilities* [Ph.D., The University of Wisconsin - Milwaukee]. https://www.proquest.com/docview/250846378/abstract/42880F0E3B474B17PQ/1

Fälth, L., Svensson, E., & Ström, A. (2020). Intensive phonological training with articulation—an intervention study to boost pupils' word decoding in grade 1. *Journal of Cognitive Education and Psychology*, *19*(2), 161–171. https://doi.org/10.1891/JCEP-D-20-00015

Fantuzzo, J. W., Davis, G. Y., & Ginsburg, M. D. (1995). Effects of parent involvement in isolation or in combination with peer tutoring on student self-concept and mathematics achievement. *Journal of Educational Psychology*, *87*(2), 272–281. https://doi.org/10.1037/0022-0663.87.2.272

Fantuzzo, J. W., King, J. A., & Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology*, *84*(3), 331–339. https://doi.org/10.1037/0022-0663.84.3.331

Fariss, L. L. (2013). *The effects of small group vocabulary instruction on second grade students' expressive vocabularies*. Virginia Polytechnic Institute and State University.

Fesler, L., Gu, A., & Chojnacki, G. (2023). Air tutors' online tutoring: Math knowledge impacts and participant math perceptions. middle years math grantee report series. In *Mathematica*. Mathematica. https://eric.ed.gov/?id=ED628638

Fien, H., Santoro, L., Baker, S. K., Park, Y., Chard, D. J., Williams, S., & Haria, P. (2011). Enhancing teacher read alouds with small-group vocabulary instruction for students with low vocabulary in first-grade classrooms. *School Psychology Review*, *40*(2), 307–318. https://doi.org/10.1080/02796015.2011.12087720

Fien, H., Smith, J. L. M., Smolkowski, K., Baker, S. K., Nelson, N. J., & Chaparro, E. (2015). An examination of the efficacy of a multitiered intervention on early reading outcomes for first grade students at risk for reading difficulties. *Journal of Learning Disabilities*, *48*(6), 602–621. https://doi.org/10.1177/0022219414521664

Finnegan, E. G. (2012). Two approaches to phonics instruction: Comparison of effects with children with significant cognitive disability. *Education and Training in Autism and Developmental Disabilities*, *47*(3), 269–279.

Fitz-Gibbon, C. T. (1975). *The role of change intervention: An experiment in cross-age tutoring*. University of California Los Angeles.

Fives, A., Kearns, N., Devaney, C., Canavan, J., Russell, D., Lyons, R., Eaton, P., & O'Brien, A. (2013). A one-to-one programme for at-risk readers delivered by older adult volunteers. *Review of Education*, *1*(3), 254–280. https://doi.org/10.1002/rev3.3016

Fraga González, G., Žarić, G., Tijms, J., Bonte, M., Blomert, L., & Van Der Molen, M. W. (2015). A randomized controlled trial on the beneficial effects of training letter-speech sound integration on reading fluency in children with dyslexia. *PLOS ONE*, *10*(12), e0143914. https://doi.org/10.1371/journal.pone.0143914

Fresko, B., & Eisenberg, T. (1985). The effect of two years of tutoring on mathematics and reading achievement. *The Journal of Experimental Education*, *53*(4), 193–201. https://doi.org/10.1080/00220973.1985.10806381

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-Assisted Learning Strategies: Making classrooms more responsive to diversity. *American Educational Research Journal*, *34*(1), 174–206. https://doi.org/10.2307/1163346

Fuchs, D., Hendricks, E., Walsh, M. E., Fuchs, L. S., Gilbert, J. K., Zhang Tracy, W., Patton, S., Davis-Perkins, N., Kim, W., Elleman, A. M., & Peng, P. (2018). Evaluating a multidimensional reading comprehension program and reconsidering the lowly reputation of tests of near-transfer. *Learning Disabilities Research & Practice*, *33*(1), 11–23. https://doi.org/10.1111/ldrp.12162

Fuchs, D., Kearns, D. M., Fuchs, L. S., Elleman, A. M., Gilbert, J. K., Patton, S., Peng, P., & Compton, D. L. (2019). Using moderator analysis to identify the first-grade children who benefit more and less from a reading comprehension program: A step toward aptitude-by-treatment interaction. *Exceptional Children*, *85*(2), 229–247. https://doi.org/10.1177/0014402918802801

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, *97*(3), 493–513. https://doi.org/10.1037/0022-0663.97.3.493

Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., & Schatschneider, C. (2008). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology*, *100*(3), 491–509. https://doi.org/10.1037/0022-0663.100.3.491

Fuchs, L. S., Fuchs, D., & Gilbert, J. K. (2019). Does the severity of students' pre-intervention math deficits affect responsiveness to generally effective first-grade intervention? *Exceptional Children*, *85*(2), 147–162. https://doi.org/10.1177/0014402918782628

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Appleton, A. C. (2002). Explicitly teaching for transfer: Effects on the mathematical problem-solving performance of students with mathematics disabilities. *Learning Disabilities Research & Practice*, *17*(2), 90–106. https://doi.org/10.1111/1540-5826.00036

Fuchs, L. S., Fuchs, D., & Karns, K. (2001). Enhancing kindergartners' mathematical development: Effects of Peer-Assisted Learning Strategies. *The Elementary School Journal*, *101*(5), 495–510. https://doi.org/10.1086/499684

Fuchs, L. S., Fuchs, D., Kazdan, S., & Allen, S. (1999). Effects of Peer-Assisted Learning Strategies in reading with and without training in elaborated help giving. *The Elementary School Journal*, *99*(3), 201–219. https://doi.org/10.1086/461923

Fuchs, L. S., Fuchs, D., Phillips, N. B., Hamlett, C. L., & Earns, K. (1995). Acquisition and transfer effects of classwide Peer-Assisted Learning Strategies in mathematics for students with varying learning histories. *School Psychology Review*, *24*(4), 604–620. https://doi.org/10.1080/02796015.1995.12085790

Fuchs, L. S., Fuchs, D., Yazdian, L., & Powell, S. R. (2002). Enhancing first-grade children's mathematical development with Peer-Assisted Learning Strategies. *School Psychology Review*, *31*(4), 569–583. https://doi.org/10.1080/02796015.2002.12086175

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. https://doi.org/10.1037/a0030127

Fuchs, L. S., Malone, A. S., Schumacher, R. F., Namkung, J., Hamlett, C. L., Jordan, N. C., Siegler, R. S., Gersten, R., & Changas, P. (2016). Supported self-explaining during fraction intervention. *Journal of Educational Psychology*, *108*(4), 493–508. https://doi.org/10.1037/edu0000073

Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., & Hamlett, C. L. (2010). The effects of strategic counting instruction, with and without deliberate practice, on number combination skill among students with mathematics difficulties. *Learning and Individual Differences*, *20*(2), 89–100. https://doi.org/10.1016/j.lindif.2009.09.003

Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., Hamlett, C. L., & Zumeta, R. O. (2009). Remediating number combination and word problem deficits among students with mathematics difficulties: A randomized control trial. *Journal of Educational Psychology*, *101*(3), 561–576. https://doi.org/10.1037/a0014701

Fuchs, L. S., Schumacher, R. F., Long, J., Namkung, J., Hamlett, C. L., Cirino, P. T., Jordan, N. C., Siegler, R., Gersten, R., & Changas, P. (2013). Improving at-risk learners' understanding of fractions. *Journal of Educational Psychology*, *105*(3), 683–700. https://doi.org/10.1037/a0032446

Fuchs, L. S., Schumacher, R. F., Long, J., Namkung, J., Malone, A. S., Wang, A., Hamlett, C. L., Jordan, N. C., Siegler, R. S., & Changas, P. (2016). Effects of intervention to improve at-risk fourth graders' understanding, calculations, and word problems with fractions. *The Elementary School Journal*, *116*(4), 625–651. https://doi.org/10.1086/686303

Fuchs, L. S., Seethaler, P. M., Powell, S. R., Fuchs, D., Hamlett, C. L., & Fletcher, J. M. (2008). Effects of preventative tutoring on the mathematical problem solving of third-grade students with math and reading difficulties. *Exceptional Children*, *74*(2), 155–173. https://doi.org/10.1177/001440290807400202

Fuchs, L. S., Seethaler, P. M., Sterba, S. K., Craddock, C., Fuchs, D., Compton, D. L., Geary, D. C., & Changas, P. (2019). *Schema-based word-problem intervention with and without embedded language comprehension instruction* (2020-13446-001). Vanderbilt University.

Gallagher, T. L. (2005). *An after-school literacy program : investigating the experiences of students with literacy difficulties, their volunteer tutors, and the tutors' transition into the teaching profession* [Brock University]. https://dr.library.brocku.ca/handle/10464/1455

Gattis, M. N., Morrow-Howell, N., McCrary, S., Lee, M., Jonson-Reid, M., McCoy, H., Tamar, K., Molina, A., & Invernizzi, M. (2010). Examining the effects of New York Experience Corps® program on young readers. *Literacy Research and Instruction*, *49*(4), 299–314. https://doi.org/10.1080/19388070903117948

Gelzheiser, L. M., Scanlon, D., Vellutino, F., Hallgren-Flynn, L., & Schatschneider, C. (2011). Effects of the interactive strategies approach—extended: A responsive and comprehensive intervention for intermediate-grade struggling readers. *The Elementary School Journal*, *112*(2), 280–306. https://doi.org/10.1086/661525

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. https://doi.org/10.3102/0002831214565787

Gilbert, J. K., Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Barquero, L. A., & Cho, E. (2013). Efficacy of a first-grade responsiveness-to-intervention prevention model for struggling readers. *Reading Research Quarterly*, *48*(2), 135–154. https://doi.org/10.1002/rrq.45

Gortazar, Lucas, Hupkau, Claudia, & Roldan, Antonio. (2023). *Online tutoring works: Experimental evidence from a program with vulnerable children* (EdWorkingPaper No. 23–743). Annenberg Institute for School Reform at Brown University. https://edworkingpapers.com/ai23-743

Greene, I., Tiernan, A. M., & Holloway, J. (2018). Cross-age peer tutoring and fluency-based instruction to achieve fluency with mathematics computation skills: A randomized controlled trial. *Journal of Behavioral Education*, *27*(2), 145–171. https://doi.org/10.1007/s10864-018-9291-1

Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *The Journal of Special Education*, *34*(2), 90–103. https://doi.org/10.1177/002246690003400204

Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *The Journal of Special Education*, *36*(2), 69–79. https://doi.org/10.1177/00224669020360020201

Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *The Journal of Special Education*, *39*(2), 66–85. https://doi.org/10.1177/00224669050390020301

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M. V., Dodge, K., Farkas, G., Fryer, R. G., Mayer, S., Pollack, H., Steinberg, L., & Stoddard, G. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, *113*(3), 738–765. https://doi.org/10.1257/aer.20210434

Harper, J. M. (2012). *The effectiveness of a group-based tutorial direct instruction program for long-term fostercare children: A randomized controlled trial.* Lakehead University.

Harper, J., & Schmidt, F. (2012). Preliminary effects of a group-based tutoring program for children in long-term foster care. *Children and Youth Services Review*, *34*(6), 1176–1182. https://doi.org/10.1016/j.childyouth.2012.01.040

Harper, J., & Schmidt, F. (2016). Effectiveness of a group-based academic tutoring program for children in foster care: A randomized controlled trial. *Children and Youth Services Review*, *67*, 238–246. https://doi.org/10.1016/j.childyouth.2016.06.009

Harty, K., Kanfush, P. M., & Riordan, M. (2019). Improving oral reading fluency and comprehension using grade-level fiction: A study of systematic reading remediation with urban youth at risk for school failure. *Reading Improvement*, *56*(2), 59–66.

Hassinger-Das, B., Jordan, N. C., & Dyson, N. (2015). Reading stories to learn math. *The Elementary School Journal*, *116*(2), 242–264. https://doi.org/10.1086/683986

Hebert, M. A., Powell, S. R., Bohaty, J., & Roehling, J. (2019). Piloting a mathematics-writing intervention with late elementary students at-risk for learning difficulties. *Learning Disabilities Research & Practice*, *34*(3), 144–157. https://doi.org/10.1111/ldrp.12202

Heller, L. R., & Fantuzzo, J. W. (1993). Reciprocal peer tutoring and parent partnership: Does parent involvement make a difference? *School Psychology Review*, *22*(3), 517–534. https://doi.org/10.1080/02796015.1993.12085670

Hickey, A. J., & Flynn, R. J. (2019). Effects of the TutorBright tutoring programme on the reading and mathematics skills of children in foster care: a randomised controlled trial. *Oxford Review of Education*, *45*(4), 519–537. https://doi.org/10.1080/03054985.2019.1607724

Horan, J. J., de Girolomo, M. A., Hill, R. L., & Shute, R. E. (1974). The effect of older-peer participant models on deficient academic performance. *Psychology in the Schools*, *11*(2), 207–212.

Hund-Reid, C., & Schneider, P. (2013). Effectiveness of phonological awareness intervention for kindergarten children with language impairment. *Canadian Journal of Speech-Language Pathology & Audiology*, *37*(1), 6–25.

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. https://doi.org/10.1080/19345747.2016.1138560

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. https://doi.org/10.1080/19345747.2020.1726539

Jacob, R., & Jacob, B. (2018). New evidence on the benefits of small group math instruction for young children. In *Center on Children and Families at Brookings* (Vol 2, 55; Evidence Speaks Reports). Center on Children and Families at Brookings.

Jayanthi, M., Gersten, R., Schumacher, R. F., Dimino, J., Smolkowski, K., & Spallone, S. (2021). Improving struggling fifth-grade students' understanding of fractions: A randomized controlled trial of an intervention that stresses both concepts and procedures. *Exceptional Children*, *88*(1), 81–100. https://doi.org/10.1177/00144029211008851

Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading*, *8*(1), 53–85. https://doi.org/10.1207/s1532799xssr0801_4

Jitendra, A. K., Dupuis, D. N., Rodriguez, M. C., Zaslofsky, A. F., Slater, S., Cozine-Corroy, K., & Church, C. (2013). A randomized controlled trial of the impact of schema-based instruction on mathematical outcomes for third-grade students with mathematics difficulties. *The Elementary School Journal*, *114*(2), 252–276. https://doi.org/10.1086/673199

Jones, C. (2018). SPARK early literacy: Testing the impact of a family-school-community partnership literacy intervention. *School Community Journal*, *28*(2), 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. https://doi.org/10.1080/10824669.2020.1809419

Jones, C. J., Christian, M., & Rice, A. (2016). The results of a randomized control trial evaluation of the SPARK literacy program. In *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED567484

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*(3), 647–660. https://doi.org/10.1037/a0029018

Kerins, M. R., Trotter, D., & Schoenbrodt, L. (2010). Effects of a tier 2 intervention on literacy measures: Lessons learned. *Child Language Teaching and Therapy*, *26*(3), 287–302. https://doi.org/10.1177/0265659009349985

King, B., & Kasim, A. (2015). *Rapid Phonics evaluation report and executive summary*. Education Endowment Foundation.

Kirkebøen, L. J., Gunnes, T., Lindenskov, L., & Rønning, M. (2021). *Didactic methods and small-group instruction for low-performing adolescents in mathematics: Results from a randomized controlled trial* (Working Paper No. 957). Discussion Papers. https://www.econstor.eu/handle/10419/249147

Kortecamp, K., & Peters, M. L. (2023). The impact of a high-dosage tutoring program on reading achievement of beginning readers: A multi-level analysis. *Journal of Education for Students Placed at Risk (JESPAR)*, 1–19. https://doi.org/10.1080/10824669.2023.2179056

Kraft, M. A., List, J. A., Livingston, J. A., & Sadoff, S. (2022). Online tutoring by college volunteers: experimental evidence from a pilot program. *AEA Papers and Proceedings*, *112*, 614–618. https://doi.org/10.1257/pandp.20221038

Lachney, R. (2002). *Adult-mediated reading instruction for third through fifth grade children with reading difficulties* [Doctor of Philosophy, Louisiana State University and Agricultural and Mechanical College]. LSU Doctoral Dissertation. https://doi.org/10.31390/gradschool_dissertations.3483

Lane, H. B., Pullen, P. C., Hudson, R. F., & Konold, T. R. (2009). Identifying essential instructional components of literacy tutoring for struggling beginning readers. *Literacy Research and Instruction*, *48*(4), 277–297. https://doi.org/10.1080/19388070902875173

Lane, K. L., Fletcher, T., Carter, E. W., Dejud, C., & Delorenzo, J. (2007). Paraprofessional-led phonological awareness training with youngsters at risk for reading and behavioral concerns. *Remedial and Special Education*, *28*(5), 266–276. https://doi.org/10.1177/07419325070280050201

Lee, C. C. (1980). The homework helper program: Volunteer service for academic and social enrichment in the elementary school. *The School Counselor*, *28*(1), 11–21.

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. https://doi.org/10.1177/0013124510381262

Lennon, J. E., & Slesinski, C. (1999). Early intervention in reading: Results of a screening and intervention program for kindergarten students. *School Psychology Review*, *28*(3), 353–364. https://doi.org/10.1080/02796015.1999.12085970

Lindo, E. J., Weiser, B., Cheatham, J. P., & Allor, J. H. (2018). Benefits of structured after-school literacy tutoring by university students for struggling elementary readers. *Reading & Writing Quarterly*, *34*(2), 117–131. https://doi.org/10.1080/10573569.2017.1357156

Lloyd, C., Edovald, T., Kiss, Z., Morris, S., Skipp, A., & Ahmed, H. (2015). Paired reading: Evaluation report and executive summary. In *Education Endowment Foundation*. Education Endowment Foundation. https://eric.ed.gov/?id=ED581127

Lloyd, C., Edovald, T., Morris, S. P., Skipp, A., KIss, Z., & Haywood, S. (2015). *Durham Shared Maths Project. Evaluation report and executive summary* [Report]. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/

Loeb, S., Novicoff, S., Pollard, C., Robinson, C., & White, S. (2023). *The effects of virtual tutoring on young readers: Results from a randomized controlled trial*. National Student Support Accelerator.

Loenen, A. (1989). The effectiveness of volunteer reading help and the nature of the reading help provided in practice. *British Educational Research Journal*, *15*(3), 297–316.

Lord, P., Bradshaw, S., Stevens, E., & Styles, B. (2015). Perry Beeches Coaching Programme: Evaluation report and executive summary. In *Education Endowment Foundation*. Education Endowment Foundation. https://eric.ed.gov/?id=ED581144

Lorenzo, S. L. (1993). *Effects of an experimental mentoring program on measures of performance of at-risk elementary students*. University of South Florida.

Mantzicopoulos, P., Morrison, D., Stone, E., & Setrakian, W. (1992). Use of the SEARCH/TEACH tutoring approach with middle-class students at risk for reading failure. *The Elementary School Journal*, *92*(5), 573–586. https://doi.org/10.1086/461707

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. https://eric.ed.gov/?id=ED560018

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. https://doi.org/10.3102/00028312211066848

Marr, M. B., Algozzine, B., Nicholson, K., & Keller Dugan, K. (2011). Building oral reading fluency with peer coaching. *Remedial and Special Education*, *32*(3), 256–264. https://doi.org/10.1177/0741932510362202

Mason, L. H., Davison, M. D., Hammer, C. S., Miller, C. A., & Glutting, J. J. (2013). Knowledge, writing, and language outcomes for a reading comprehension and writing intervention. *Reading and Writing*, *26*(7), 1133–1158. https://doi.org/10.1007/s11145-012-9409-0

Mathes, P. G., & Babyak, A. E. (2001). The effects of Peer-Assisted Literacy Strategies for first-grade readers with and without additional mini-skills lessons. *Learning Disabilities Research & Practice*, *16*(1), 28–44. https://doi.org/10.1111/0938-8982.00004

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, *40*(2), 148–182. https://doi.org/10.1598/RRQ.40.2.2

Mathes, P. G., Torgesen, J. K., Clancy-Menchetti, J., Santi, K., Nicholas, K., Robinson, C., & Grek, M. (2003). A comparison of teacher-directed versus peer-assisted instruction to struggling first-grade readers. *The Elementary School Journal*, *103*(5), 459–479. https://doi.org/10.1086/499735

Mattera, S., Jacob, R., & Morris, P. (2018). *Strengthening children's math skills with enhanced instruction: the impacts of Making Pre-K Count and High 5s on kindergarten outcomes* (SSRN Scholarly Paper No. 3167484). https://papers.ssrn.com/abstract=3167484

Mattera, S. K., Jacob, R., MacDowell, C., & Morris, P. A. (2021). Long-term effects of enhanced early childhood math instruction: the impacts of Making Pre-K Count and High 5s on third-grade outcomes. In *MDRC*. MDRC. https://eric.ed.gov/?id=ED616510

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., & Clague, L. (2014). Textnow transition programme: Evaluation report and executive summary. In *Education Endowment Foundation*. Education Endowment Foundation. https://eric.ed.gov/?id=ED581131

May, H., Goldsworthy, H., Armijo, M., Gray, A. M., Sirinides, P., Blalock, T. J., Anderson-Clark, H., Schiera, A. J., Blackman, H., Gillespie, J., & Sam, C. (2014). *Evaluation of the i3 scale-up of Reading Recovery | Year two report, 2012-13* [CPRE Research Reports]. University of Pennsylvania Consortium for Policy Research in Education.

May, H., Gray, A., Sirinides, P., Goldsworthy, H., Armijo, M., Sam, C., Gillespie, J. N., & Tognatta, N. (2015). Year one results from the multisite randomized evaluation of the i3 scale-up of reading recovery. *American Educational Research Journal*, *52*(3), 547–581. https://doi.org/10.3102/0002831214565788

Mayfield, L. (2000). *The effects of structured one-on-one tutoring in sight word recognition of first-grade students at-risk for reading failure* [Louisiana Tech University]. https://digitalcommons.latech.edu/dissertations/185

McCleary, E. K. (1971). Report of results of Tutorial Reading Project. *The Reading Teacher*, *24*(6), 556–559.

McKinney, A. D. (1995). *The effects of an after school tutorial and enrichment program on the academic achievement and self-concept of below grade level first and second-grade students*. The University of Mississippi.

Menesses, K. F., & Gresham, F. M. (2009). Relative efficacy of reciprocal and nonreciprocal peer tutoring for students at-risk for academic failure. *School Psychology Quarterly*, *24*(4), 266–275. https://doi.org/10.1037/a0018174

Merrell, C., & Kasim, A. (2015). Butterfly Phonics: Evaluation report and executive summary. In *Education Endowment Foundation*. Education Endowment Foundation. https://eric.ed.gov/?id=ED581118

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.

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. https://doi.org/10.1177/1476718X11407989

Mooney, P. J. (2003). *An investigation of the effects of a comprehensive reading intervention on the beginning reading skills of first graders at risk for emotional and behavioral disorders* [Ph.D., The University of Nebraska - Lincoln]. https://www.proquest.com/docview/305310607/abstract/AE7141CF198343FDPQ/1

Moran, A. S., Swanson, H. L., Gerber, M. M., & Fung, W. (2014). The effects of paraphrasing interventions on problem-solving accuracy for children at risk for math disabilities. *Learning Disabilities Research & Practice*, *29*(3), 97–105. https://doi.org/10.1111/ldrp.12035

Morris, D., Shaw, B., & Perney, J. (1990). Helping low readers in grades 2 and 3: An after-school volunteer tutoring program. *The Elementary School Journal*, *91*(2), 133–150. https://doi.org/10.1086/461642

Moser Opitz, E., Freesemann, O., Prediger, S., Grob, U., Matull, I., & Hußmann, S. (2017). Remediation for students with mathematics difficulties: An intervention study in middle schools. *Journal of Learning Disabilities*, *50*(6), 724–736. https://doi.org/10.1177/0022219416668323

Murdaugh, D. L., Maximo, J. O., Cordes, C. E., O'Kelley, S. E., & Kana, R. K. (2017). From word reading to multisentence comprehension: Improvements in brain activity in children with autism after reading intervention. *NeuroImage: Clinical*, *16*, 303–312. https://doi.org/10.1016/j.nicl.2017.08.012

Nelson, R. J., Benner, G. J., & Gonzalez, J. (2005). An investigation of the effects of a prereading intervention on the early literacy skills of children at risk of emotional disturbance and reading problems. *Journal of Emotional and Behavioral Disorders*, *13*(1), 3–12. https://doi.org/10.1177/10634266050130010101

Nielsen, D. C., & Friesen, L. D. (2012). A study of the effectiveness of a small-group intervention on the vocabulary and narrative development of at-risk kindergarten children. *Reading Psychology*, *33*(3), 269–299. https://doi.org/10.1080/02702711.2010.508671

Nielson, B. B. (1991). *Effects of parent and volunteer tutoring on reading achievement of third grade at-risk students*. Brigham Young University.

Nunes, T., Barros, R., Evangelou, M., Strand, S., Mathers, S., & Sanders-Ellis, D. (2018). *1stClass@Number Evaluation report and executive summary*. Education Endowment Foundation.

Nunes, T., Bryant, P., & Olsson, J. (2003). Learning morphological and phonological spelling rules: An intervention study. *Scientific Studies of Reading*, *7*(3), 289–307. https://doi.org/10.1207/S1532799XSSR0703_6

Oakland, T., & Williams, F. C. (1975). An evaluation of two methods of peer tutoring. *Psychology in the Schools*, *12*(2), 166–171.

O'Connor, R. E., Bell, K. M., Harty, K. R., Larkin, L. K., Sackor, S. M., & Zigmond, N. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology*, *94*(3), 474–485. https://doi.org/10.1037/0022-0663.94.3.474

O'Connor, R. E., Bocian, K., Beebe-Frankenberger, M., & Linklater, D. L. (2010). Responsiveness of students with language difficulties to early intervention in reading. *The Journal of Special Education*, *43*(4), 220–235. https://doi.org/10.1177/0022466908317789

O'Connor, R. E., Swanson, H. L., & Geraghty, C. (2010). Improvement in reading rate under independent and difficult text levels: Influences on word and comprehension skills. *Journal of Educational Psychology*, *102*(1), 1–19. https://doi.org/10.1037/a0017488

Oostdam, R., Blok, H., & Boendermaker, C. (2015). Effects of individualised and small-group guided oral reading interventions on reading skills and reading attitude of poor readers in grades 2–4. *Research Papers in Education*, *30*(4), 427–450. https://doi.org/10.1080/02671522.2014.953195

Paramore, B., Plantec, P., & Hospodar, J. (1973). *Project Upswing after two years: An evaluation* (TR-801). Operations Research, Inc. https://eric.ed.gov/?id=ED099820

Parham, J. W. (1993). *An analysis of the effects of tutoring on seventh-grade students engaged in the mastery of pre-algebra concepts*. University of Maryland College Park.

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. https://doi.org/10.1080/19345747.2019.1571653

Patton, S. A., Fuchs, D., Hendricks, E. L., Pennell, A. J., Walsh, M. E., Fuchs, L. S., Tracy, W. Z., & Haga, L. Y. (2022). An experimental study to strengthen students' comprehension of informational texts: Is teaching for transfer important? *Learning Disabilities Research & Practice*, *37*(2), 124–139. https://doi.org/10.1111/ldrp.12276

Peng, P., & Fuchs, D. (2017). A randomized control trial of working memory training with and without strategy instruction: Effects on young children's working memory and comprehension. *Journal of Learning Disabilities*, *50*(1), 62–80. https://doi.org/10.1177/0022219415594609

Pericola Case, L., Speece, D. L., Silverman, R., Ritchey, K. D., Schatschneider, C., Cooper, D. H., Montanaro, E., & Jacobs, D. (2010). Validation of a supplemental reading intervention for first-grade children. *Journal of Learning Disabilities*, *43*(5), 402–417. https://doi.org/10.1177/0022219409355475

Pinnell, G. S., Lyons, C. A., DeFord, D. E., Bryk, A. S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly*, *29*(1), 9–39. https://doi.org/10.2307/747736

Powell, S. R., Berry, K. A., Acunto, A. N., Fall, A.-M., & Roberts, G. (2022). Applying an individual word-problem intervention to a small-group setting: A pilot study's evidence of improved word-problem performance for students experiencing mathematics difficulty. *Journal of Learning Disabilities*, *55*(5), 359–374. https://doi.org/10.1177/00222194211047635

Powell, S. R., & Driver, M. K. (2015). The influence of mathematics vocabulary instruction embedded within addition tutoring for first-grade students with mathematics difficulty. *Learning Disability Quarterly*, *38*(4), 221–233. https://doi.org/10.1177/0731948714564574

Powell, S. R., Driver, M. K., & Julian, T. E. (2015). The effect of tutoring with nonstandard equations for students with mathematics difficulty. *Journal of Learning Disabilities*, *48*(5), 523–534. https://doi.org/10.1177/0022219413512613

Powell, S. R., & Fuchs, L. S. (2010). Contribution of equal-sign instruction beyond word-problem tutoring for third-grade students with mathematics difficulty. *Journal of Educational Psychology*, *102*(2), 381–394. https://doi.org/10.1037/a0018447

Powell, S. R., Fuchs, L. S., Fuchs, D., Cirino, P. T., & Fletcher, J. M. (2009). Effects of fact retrieval tutoring on third-grade students with math difficulties with and without reading difficulties. *Learning Disabilities Research & Practice*, *24*(1), 1–11. https://doi.org/10.1111/j.1540-5826.2008.01272.x

Powell, S. R., Stevens, E. A., & Berry, K. A. (2019). Effects of a word-problem intervention on word-problem language features for third-grade students with mathematics difficulty. *Learning Disabilities: A Multidisciplinary Journal*, *24*(2), 1–14. https://doi.org/10.18666/LDMJ-2019-V24-I2-9835

Pullen, P. C., & Lane, H. B. (2014). Teacher-directed decoding practice with manipulative letters and word reading skill development of struggling first grade students. *Exceptionality*, *22*(1), 1–16. https://doi.org/10.1080/09362835.2014.865952

Pullen, P. C., Lane, H. B., & Monaghan, M. C. (2004). Effects of a volunteer tutoring model on the early literacy development of struggling first grade students. *Reading Research and Instruction*, *43*(4), 21–40. https://doi.org/10.1080/19388070409558415

Ransford-Kaldon, C. R., Flynt, E. S., Ross, C. L., Franceschini, L., Zoblotsky, T., Huang, Y., & Gallagher, B. (2010). Implementation of effective intervention: An empirical study to evaluate the efficacy of Fountas & Pinnell's Leveled Literacy Intervention system (LLI). 2009-2010. *Center for Research in Educational Policy (CREP)*. https://eric.ed.gov/?id=ed544374

Ransford-Kaldon, C. R., Ross, C. L., Lee, C. C., Flynt, E. S., Franceschini, L., & Zoblotsky, T. (2013). *Efficacy of the leveled literacy intervention system for K-2 urban students: An empirical evaluation of LLI in Denver Public Schools*. Center for Research in Education Policy.

Rashotte, C. A., MacPhee, K., & Torgesen, J. K. (2001). The effectiveness of a group reading instruction program with poor readers in multiple grades. *Learning Disability Quarterly*, *24*(2), 119–134. https://doi.org/10.2307/1511068

Rebok, G. W., Carlson, M. C., Glass, T. A., McGill, S., Hill, J., Wasik, B. A., Ialongo, N., Frick, K. D., Fried, L. P., & 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. https://doi.org/10.1093/jurban/jth095
Reynolds, D. (2021). Scaffolding the academic language of complex text: an intervention for late secondary students. *Journal of Research in Reading*, *44*(3), 508–528. https://doi.org/10.1111/1467-9817.12353
Rimm-Kaufman, S. E., Kagan, J., & Byers, H. (1998). The effectiveness of adult volunteer tutoring on reading among "at risk" first grade children. *Reading Research and Instruction*, *38*(2), 143–152. https://doi.org/10.1080/19388079909558284
Ritchey, K. D., Silverman, R. D., Montanaro, E. A., Speece, D. L., & Schatschneider, C. (2012). Effects of a tier 2 supplemental reading intervention for at-risk fourth-grade students. *Exceptional Children*, *78*(3), 318–334. https://doi.org/10.1177/001440291207800304
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*, *3*(2), 4–16. https://doi.org/10.1108/17466660200800008
Roach, J. C., Paolucci-Whitcomb, P., Meyers, H. W., & Duncan, D. A. (1983). The comparative effects of peer tutoring in math by and for secondary special needs students. *Pointer*, *27*(4), 20–24.
Roberts, G. J., Capin, P., Roberts, G., Miciak, J., Quinn, J. M., & Vaughn, S. (2018). Examining the effects of afterschool reading interventions for upper elementary struggling readers. *Remedial and Special Education*, *39*(3), 131–143. https://doi.org/10.1177/0741932517750818
Rogers, J. M. (1979). *The effects of tutoring by sixth graders on the reading performance of first graders.* [Educat.D., University of San Francisco]. https://www.proquest.com/docview/303035491/citation/83642FE214984B85PQ/1
Rolfhus, E., Gersten, R., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2012). An evaluation of "Number Rockets": A tier-2 intervention for grade 1 students at risk for difficulties in mathematics. Final report. In *National Center for Education Evaluation and Regional Assistance* (No. 2012–4007). National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED529429
Roschelle, J., Cheng, B. H., Hodkowski, N., Neisler, J., & Haldar, L. (2020). Evaluation of an online tutoring program in elementary mathematics. In *Online Submission*. Digital Promise. https://eric.ed.gov/?id=ED604743
Rosenshine, B., Mattleman, M., & Rosenshine, B. (1971). *Remediation in reading for fourth graders: A project report for 1969-70*. Philadelphia School District. https://eric.ed.gov/?id=ED053879
Rutt, S., Kettlewell, K., & Bernardinelli, D. (2015). Catch Up® Literacy: Evaluation report and executive summary. In *National Foundation for Educational Research*. National Foundation for Educational Research. https://eric.ed.gov/?id=ED558735
Ryder, J. F., Tunmer, W. E., & Greaney, K. T. (2008). Explicit instruction in phonemic awareness and phonemically based decoding skills as an intervention strategy for struggling readers in whole language classrooms. *Reading and Writing*, *21*(4), 349–369. https://doi.org/10.1007/s11145-007-9080-z
Scanlon, D. M., Vellutino, F. R., Small, S. G., Fanuele, D. P., & Sweeney, J. M. (2005). Severe reading difficulties—can they be prevented? A comparison of prevention and intervention approaches. *Exceptionality*, *13*(4), 209–227. https://doi.org/10.1207/s15327035ex1304_3

Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the Reading Recovery early intervention. *Journal of Educational Psychology*, *97*(2), 257–267. https://doi.org/10.1037/0022-0663.97.2.257

See, B. H., Morris, R., Gorard, S., & Siddiqui, N. (2019). Evaluation of the impact of Maths Counts delivered by teaching assistants on primary school pupils' attainment in maths. *Educational Research and Evaluation*, *25*(3–4), 203–224. https://doi.org/10.1080/13803611.2019.1686031

Sharma, M., Purdy, S. C., & Kelly, A. S. (2012). A randomized control trial of interventions in school-aged children with auditory processing disorders. *International Journal of Audiology*, *51*(7), 506–518. https://doi.org/10.3109/14992027.2012.670272

Sibieta, L., Kotecha, M., & Skipp, A. (2016). *Nuffield Early Language Intervention: Evaluation report and executive summary*. Education Endowment Foundation. https://eric.ed.gov/?id=ED581138

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. https://doi.org/10.3102/0162373718764828

Smith, S. B. (1996). *An examination of the efficacy and the efficiency of phonological awareness instruction for prereaders at-risk of reading failure. Final report* [University of Oregon]. https://eric.ed.gov/?id=ED421805

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. https://doi.org/10.3102/0002831212469045

Solari, E. J., Denton, C. A., Petscher, Y., & Haring, C. (2018). Examining the effects and feasibility of a teacher-implemented tier 1 and tier 2 intervention in word reading, fluency, and comprehension. *Journal of Research on Educational Effectiveness*, *11*(2), 163–191. https://doi.org/10.1080/19345747.2017.1375582

Solís, M., Scammacca, N., Barth, A. E., & Roberts, G. J. (2017). Text-based vocabulary intervention training study: Supporting fourth graders with low reading comprehension and learning disabilities. *Learning Disabilities (Weston, Mass.)*, *15*(1), 103–115.

Solis, M., Vaughn, S., Stillman-Spisak, S. J., & Cho, E. (2018). Effects of reading comprehension and vocabulary intervention on comprehension-related outcomes for ninth graders with low reading comprehension. *Reading & Writing Quarterly*, *34*(6), 537–553. https://doi.org/10.1080/10573569.2018.1499059

Swanson, H. L. (2015). Cognitive strategy interventions improve word problem solving and working memory in children with math disabilities. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01099

Swanson, H. L., Kong, J. E., Moran, A. S., & Orosco, M. J. (2019). Paraphrasing interventions and problem-solving accuracy: Do generative procedures help english language learners with math difficulties? *Learning Disabilities Research & Practice*, *34*(2), 68–84. https://doi.org/10.1111/ldrp.12194

Swanson, H. L., Moran, A., Lussier, C., & Fung, W. (2014). The effect of explicit and direct generative strategy training and working memory on word problem-solving accuracy in children at risk for math difficulties. *Learning Disability Quarterly*, *37*(2), 111–123. https://doi.org/10.1177/0731948713507264

Swanson, H. L., Orosco, M. J., & Lussier, C. M. (2014). The effects of mathematics strategy instruction for children with serious problem-solving difficulties. *Exceptional Children*, *80*(2), 149–168. https://doi.org/10.1177/001440291408000202

Thurmann-Moe, A. C., Melby-Lervåg, M., & Lervåg, A. (2022). Effects of articulatory consciousness training in first graders with a reading delay: A randomised control trial. *Scandinavian Journal of Educational Research*, *66*(3), 473–490. https://doi.org/10.1080/00313831.2020.1869823

Thurston, A., Cockerill, M., & Chiang, T.-H. (2021). Assessing the differential effects of peer tutoring for tutors and tutees. *Education Sciences*, *11*(3), 97. https://doi.org/10.3390/educsci11030097

Thurston, A., Cockerill, M., & Craig, N. (2019). Using cooperative learning to close the reading attainment gap for students with low literacy levels for Grade 8/Year 9 students. *International Journal of Educational Research*, *94*, 1–10. https://doi.org/10.1016/j.ijer.2019.02.016

Tolan, P., Gorman-Smith, D., & Henry, D. (2004). Supporting families in a high-risk setting: Proximal effects of the SAFEChildren preventive intervention. *Journal of Consulting and Clinical Psychology*, *72*(5), 855–869. https://doi.org/10.1037/0022-006X.72.5.855

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., Jefferson, L., Mitchell, N., & Torgerson, D. (2016). Affordable online maths tuition: Evaluation report and executive summary. In *Education Endowment Foundation*. Education Endowment Foundation. https://eric.ed.gov/?id=ED581116

Torgerson, C., Bell, K., Coleman, E., Elliot, L., Fairhurst, C., Gascione, L., Hewitt, C. E., & Torgerson, D. (2018). *Tutor Trust: Affordable primary tuition evaluation report and executive summary November 2018*. Education Endowment Foundation.

Torgerson, C., Wiggins, A., Torgerson, D., Ainsworth, H., & Hewitt, C. (2013). *Every Child Counts* : testing policy effectiveness using a randomised controlled trial, designed, conducted and reported to CONSORT standards. *Research in Mathematics Education*, *15*(2), 141–153. https://doi.org/10.1080/14794802.2013.797746

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Herron, J., & Lindamood, P. (2010). Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: Outcomes from two instructional approaches. *Annals of Dyslexia*, *60*(1), 40–56. https://doi.org/10.1007/s11881-009-0032-y

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. https://doi.org/10.1037/0022-0663.91.4.579

Torgesen, J., Myers, D., Schirm, A., Stuart, E. A., Vartivarian, S., Mansfield, W., Stancavage, F., Durno, D., Javorsky, R., & Haan, C. (2006). *Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers* (0f5955350f2f4dc6bd83a19896ba3472; Mathematica Policy Research Reports). Mathematica Policy Research. https://ideas.repec.org//p/mpr/mprres/0f5955350f2f4dc6bd83a19896ba3472.html

Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., Stancavage, F., Durno, D., Javorsky, R., & Haan, C. (2007). National assessment of Title I. Final report. Volume II: Closing the reading gap--findings from a randomized trial of four reading interventions for striving readers. In *National Center for Education Evaluation and*

Regional Assistance (No. 2008–4013). National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED499018

Toste, J. R., Capin, P., Vaughn, S., Roberts, G. J., & Kearns, D. M. (2017). Multisyllabic word-reading instruction with and without motivational beliefs training for struggling readers in the upper elementary grades: A pilot investigation. *The Elementary School Journal*, *117*(4), 593–615. https://doi.org/10.1086/691684

Toste, J. R., Capin, P., Williams, K. J., Cho, E., & Vaughn, S. (2019). Replication of an experimental study investigating the efficacy of a multisyllabic word reading intervention with and without motivational beliefs training for struggling readers. *Journal of Learning Disabilities*, *52*(1), 45–58. https://doi.org/10.1177/0022219418775114

Tournaki, N. (2003). The differential effects of teaching addition through strategy instruction versus drill and practice to students with and without learning disabilities. *Journal of Learning Disabilities*, *36*(5), 449–458. https://doi.org/10.1177/00222194030360050601

Vadasy, P. F., Jenkins, J. R., Antil, L. R., Wayne, S. K., & O'Connor, R. E. (1997a). Community-based early reading intervention for at-risk first graders. *Learning Disabilities Research & Practice*, *12*(1), 29–39.

Vadasy, P. F., Jenkins, J. R., Antil, L. R., Wayne, S. K., & O'Connor, R. E. (1997b). The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers. *Learning Disability Quarterly*, *20*(2), 126–139. https://doi.org/10.2307/1511219

Vadasy, P. F., Jenkins, J. R., & Pool, K. (2000). Effects of tutoring in phonological and early reading skills on students at risk for reading disabilities. *Journal of Learning Disabilities*, *33*(6), 579–590. https://doi.org/10.1177/002221940003300606

Vadasy, P. F., & Sanders, E. A. (2008a). Repeated reading intervention: Outcomes and interactions with readers' skills and classroom instruction. *Journal of Educational Psychology*, *100*(2), 272–290. https://doi.org/10.1037/0022-0663.100.2.272

Vadasy, P. F., & Sanders, E. A. (2008b). Benefits of repeated reading intervention for low-achieving fourth- and fifth-grade students. *Remedial and Special Education*, *29*(4), 235–249. https://doi.org/10.1177/0741932507312013

Vadasy, P. F., & Sanders, E. A. (2008c). Code-oriented instruction for kindergarten students at risk for reading difficulties: a replication and comparison of instructional groupings. *Reading and Writing*, *21*(9), 929–963. https://doi.org/10.1007/s11145-008-9119-9

Vadasy, P. F., & Sanders, E. A. (2009). Supplemental fluency intervention and determinants of reading outcomes. *Scientific Studies of Reading*, *13*(5), 383–425. https://doi.org/10.1080/10888430903162894

Vadasy, P. F., & Sanders, E. A. (2010). Efficacy of supplemental phonics-based instruction for low-skilled kindergarteners in the context of language minority status and classroom phonics instruction. *Journal of Educational Psychology*, *102*(4), 786–803. https://doi.org/10.1037/a0019639

Vadasy, P. F., & Sanders, E. A. (2011). Efficacy of supplemental phonics-based instruction for low-skilled first graders: How language minority status and pretest characteristics moderate treatment response. *Scientific Studies of Reading*, *15*(6), 471–497. https://doi.org/10.1080/10888438.2010.501091

Vadasy, P. F., & Sanders, E. A. (2012). Two-year follow-up of a kindergarten phonics intervention for English learners and native English speakers: Contextualizing treatment impacts by classroom literacy instruction. *Journal of Educational Psychology*, *104*(4), 987–1005. https://doi.org/10.1037/a0028163

Vadasy, P. F., & Sanders, E. A. (2013). Two-year follow-up of a code-oriented intervention for lower-skilled first-graders: The influence of language status and word reading skills on third-grade literacy outcomes. *Reading and Writing*, *26*(6), 821–843. https://doi.org/10.1007/s11145-012-9393-4

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006a). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology*, *98*(3), 508–528. https://doi.org/10.1037/0022-0663.98.3.508

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006b). Paraeducator-supplemented instruction in structural analysis with text reading practice for second and third graders at risk for reading problems. *Remedial and Special Education*, *27*(6), 365–378. https://doi.org/10.1177/07419325060270060601

Vadasy, P. F., Sanders, E. A., & Tudor, S. (2007). Effectiveness of paraeducator-supplemented individual instruction: Beyond basic decoding skills. *Journal of Learning Disabilities*, *40*(6), 508–525. https://doi.org/10.1177/00222194070400060301

Van Luit, J. E. H., & Naglieri, J. A. (1999). Effectiveness of the MASTER Program for teaching special children multiplication and division. *Journal of Learning Disabilities*, *32*(2), 98–107. https://doi.org/10.1177/002221949903200201

Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., Hagan, E. C., Pollard-Durodola, S. D., Fletcher, J. M., & Francis, D. J. (2006). Effectiveness of a Spanish intervention and an English intervention for English-language learners at risk for reading problems. *American Educational Research Journal*, *43*(3), 449–487. https://doi.org/10.3102/00028312043003449

Vaughn, S., Linan-Thompson, S., Mathes, P. G., Cirino, P. T., Carlson, C. D., Pollard-Durodola, S. D., Cardenas-Hagan, E., & Francis, D. J. (2006). Effectiveness of Spanish intervention for first-grade English language learners at risk for reading difficulties. *Journal of Learning Disabilities*, *39*(1), 56–73. https://doi.org/10.1177/00222194060390010601

Vaughn, S., Mathes, P., Linan-Thompson, S., Cirino, P., Carlson, C., Pollard-Durodola, S., Cardenas-Hagan, E., & Francis, D. (2006). Effectiveness of an English intervention for first-grade English language learners at risk for reading problems. *The Elementary School Journal*, *107*(2), 153–180. https://doi.org/10.1086/510653

Vaughn, S., Roberts, G. J., Miciak, J., Taylor, P., & Fletcher, J. M. (2019). Efficacy of a word- and text-based intervention for students with significant reading difficulties. *Journal of Learning Disabilities*, *52*(1), 31–44. https://doi.org/10.1177/0022219418775113

Vaughn, S., Wanzek, J., Wexler, J., Barth, A., Cirino, P. T., Fletcher, J., Romain, M., Denton, C. A., Roberts, G., & Francis, D. (2010). The relative effects of group size on reading progress of older students with reading difficulties. *Reading and Writing*, *23*(8), 931–956. https://doi.org/10.1007/s11145-009-9183-9

Vernon-Feagans, L., Kainz, K., Hedrick, A., Ginsberg, M., & Amendum, S. (2010). The targeted reading intervention: A classroom teacher professional development program to promote effective teaching for struggling readers in kindergarten and first grade. In *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED512686

Villiger, C., Hauri, S., Tettenborn, A., Hartmann, E., Näpflin, C., Hugener, I., & Niggli, A. (2019). Effectiveness of an extracurricular program for struggling readers: A comparative study with parent tutors and volunteer tutors. *Learning and Instruction*, *60*, 54–65. https://doi.org/10.1016/j.learninstruc.2018.11.004

Vousden, J. I., Cunningham, A. J., Johnson, H., Waldron, S., Ammi, S., Pillinger, C., Savage, R., & Wood, C. (2022). Decoding and comprehension skills mediate the link between a small-group reading programme and English national literacy assessments. *British Journal of Educational Psychology*, *92*(1), 105–130. https://doi.org/10.1111/bjep.12441

Walsh, M. E. (2020). *Developing an informational text comprehension intervention for struggling readers in third grade* [Ph.D., Vanderbilt University]. https://www.proquest.com/docview/2440672823/abstract/3EF0E56DA75C4F62PQ/1

Wanzek, J., Otaiba, S. A., Schatschneider, C., Donegan, R. E., Rivas, B., Jones, F., & Petscher, Y. (2020). Intensive intervention for upper elementary students with severe reading comprehension difficulties. *Journal of Research on Educational Effectiveness*, *13*(3), 408–429. https://doi.org/10.1080/19345747.2019.1710886

Wanzek, J., Petscher, Y., Otaiba, S. A., Rivas, B. K., Jones, F. G., Kent, S. C., Schatschneider, C., & Mehta, P. (2017). Effects of a year long supplemental reading intervention for students with reading difficulties in fourth grade. *Journal of Educational Psychology*, *109*(8), 1103–1119. https://doi.org/10.1037/edu0000184

Wanzek, J., & Roberts, G. (2012). Reading interventions with varying instructional emphases for fourth graders with reading difficulties. *Learning Disability Quarterly*, *35*(2), 90–101. https://doi.org/10.1177/0731948711434047

Wanzek, J., & Vaughn, S. (2008). Response to varying amounts of time in reading intervention for students with low response to intervention. *Journal of Learning Disabilities*, *41*(2), 126–142. https://doi.org/10.1177/0022219407313426

Wexler, J., Vaughn, S., Roberts, G., & Denton, C. A. (2010). The efficacy of Repeated Reading and Wide Reading Practice for high school students with severe reading disabilities. *Learning Disabilities Research & Practice*, *25*(1), 2–10. https://doi.org/10.1111/j.1540-5826.2009.00296.x

White, J. R. (2000). *A study of the effectiveness of using volunteer tutors in a corrective reading program*. Old Dominion University.

White, M. M. (2012). *Repeated reading of core text: A study of reading fluency among below-level readers* [Ph.D., Capella University]. https://www.proquest.com/docview/945731824/abstract/542E539E61834B69PQ/1

Wilkerson, S. B. (2008). *A study of Pearson's My Sidewalks Program: Final report*. Magnolia Consulting.

Wolff, U. (2011). Effects of a randomised reading intervention study: An application of structural equation modelling. *Dyslexia*, *17*(4), 295–311. https://doi.org/10.1002/dys.438

Woo, D. G. (2005). *America Reads: The effects of a federal work -study tutoring program on literacy achievement and attitudes of teachers, tutors, and children* [Ed.D., Rutgers The State University of New Jersey, School of Graduate Studies]. https://www.proquest.com/docview/305396827/abstract/58FC4F3254EF4E39PQ/1

Wright, H., Dorsett, R., Anders, J., Buzzeo, J., Runge, J., & Sanders, M. (2019). Improving Working Memory: Evaluation report and executive summary. In *The Education Endowment Foundation (EEF): London, UK.* [Report]. The Education Endowment Foundation (EEF). https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Improving_Working_Memory_Report_final.pdf

Young, C., Pearce, D., Gomez, J., Christensen, R., Pletcher, B., & Fleming, K. (2018). Read Two Impress and the Neurological Impress Method: Effects on elementary students' reading fluency, comprehension, and attitude. *The Journal of Educational Research*, *111*(6), 657–665. https://doi.org/10.1080/00220671.2017.1393650

Zinn, A., & Courtney, M. E. (2014). Context matters: Experimental evaluation of home-based tutoring for youth in foster care. *Children and Youth Services Review*, *47*, 198–204. https://doi.org/10.1016/j.childyouth.2014.08.017

Zvoch, K. (2019). Investigation of the long term effect of a summer literacy program on student reading performance. *Studies in Educational Evaluation*, *62*, 111–117. https://doi.org/10.1016/j.stueduc.2019.05.005

Appendix B. Additional Tables and Figures

Table B1. Trim and fill estimates, stacked subjects

Effect-level estimates		Study-by-subject-level estimates	
Observed	Observed and imputed	Observed	Observed and imputed
(1)	(2)	(3)	(4)
Panel A. Full sample			
0.444***	0.444***	0.348***	0.348***
(0.014)	(0.014)	(0.023)	(0.023)
2,223	2,223	277	277
Panel B. Studies with 0-99 treated students			
0.550***	0.550***	0.458***	0.458***
(0.021)	(0.021)	(0.039)	(0.039)
1,403	1,403	166	166
Panel C. Studies with ≥100 students			
0.289***	0.289***	0.239***	0.239***
(0.013)	(0.013)	(0.022)	(0.022)
820	820	111	111

Notes: *** *p*<0.10; ** *p*<0.05, * *p*<0.01. All estimates stack math and reading. Columns (1) and (2) report trim and fill estimate results on effect size level observations, while columns (3) and (4) report the same for observations collapsed to the study by subject area level. While Panel A conducts this estimation on the full sample, Panels B and C subdivide the sample into studies of programs serving less than and at least 100 tutored students, respectively. Columns (2) and (4) impute values if there are imbalances in the distribution of effects reported in columns (1) and (3), respectively. Note that if the distribution of effects was imbalanced around the average pooled estimate, the samples in columns (2) and (4) would include additional, imputed values that fill in the "gaps" of those distributions; observing no differences in these samples is evidence results are not meaningfully impacted by publication bias.

Table B2. Estimates pooled by publication type

No sample size restriction	0-99 treated students	100-399 treated students	400-999 treated students	≥1000 treated students
(1)	(2)	(3)	(4)	(5)
Panel A. Subsample of studies published in peer-reviewed journals				
0.454***	0.543***	0.340***	0.279***	0.325*
(0.040)	(0.064)	(0.029)	(0.048)	(0.193)
1940	1288	570	67	15
Panel B. Subsample of studies not published in peer-reviewed journals				
0.220***	0.572***	0.150***	0.207***	0.085
(0.056)	(0.133)	(0.029)	(0.052)	(0.064)
283	115	70	45	53

Notes: *** *p*<0.10; ** *p*<0.05, * *p*<0.01. Column (1) present the pooled meta-analytic average effect size estimate for the subsample of effects described by each panel. This estimate is disaggregated across treated student sample size in columns (2) through (5). Each panel isolates a single feature of tutoring program design that has been associated with elevated impacts in prior research.

Table B3. Estimates pooled by best practices, stacked subjects

No sample size restriction (1)	0-99 treated students (2)	100-399 treated students (3)	400-999 treated students (4)	≥1000 treated students (5)
Panel A. Subsample of tutoring programs conducted in-person				
0.438***	0.545***	0.320***	0.253***	0.276**
(0.036)	(0.060)	(0.026)	(0.037)	(0.118)
2,164	1,396	630	112	26
Panel B. Subsample of tutoring programs conducted at school				
0.404***	0.504***	0.318***	0.260***	0.138
(0.030)	(0.045)	(0.030)	(0.038)	(0.103)
1,885	1,236	473	108	68
Panel C. Subsample of tutoring programs conducted during the school day				
0.388***	0.473***	0.324***	0.272***	0.147
(0.027)	(0.039)	(0.028)	(0.040)	(0.125)
1,748	1,086	496	103	63
Panel D. Subsample of tutoring programs with student-tutor ratios of no more than 3:1				
0.400***	0.504***	0.327***	0.245***	0.147
(0.030)	(0.047)	(0.028)	(0.042)	(0.129)
1,665	999	531	73	62
Panel E. Subsample of tutoring programs delivering ≥15 total hours of tutoring				
0.423***	0.523***	0.332***	0.254***	0.306***
(0.039)	(0.070)	(0.032)	(0.042)	(0.115)
1,572	1,003	447	92	30
Panel F. Subsample of tutoring programs using curricula for lessons				
0.410***	0.518***	0.330***	0.260***	0.055***
(0.036)	(0.059)	(0.027)	(0.038)	(0.008)
2,011	1,245	602	109	55
Panel G. Subsample of tutoring programs with ≥3 sessions per week				
0.411***	0.509***	0.348***	0.260***	0.150
(0.030)	(0.049)	(0.026)	(0.045)	(0.135)
1,805	1,086	573	85	61

Notes: *** p<0.10; ** p<0.05, * p<0.01. Column (1) present the pooled meta-analytic average effect size estimate for the subsample of effects described by each panel. This estimate is disaggregated across treated student sample size in columns (2) through (5). Each panel isolates a single feature of tutoring program design that has been associated with elevated impacts in prior research.