NBER WORKING PAPER SERIES

THE IMPRESSIVE EFFECTS OF TUTORING ON PREK-12 LEARNING: A SYSTEMATIC REVIEW AND META-ANALYSIS OF THE EXPERIMENTAL EVIDENCE

Andre Nickow
Philip Oreopoulos
Vincent Quan

Working Paper 27476
http://www.nber.org/papers/w27476

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2020

We would like to thank Kimberly Dadisman for her invaluable work in moving the project forward, Bradley Clark for his excellent research assistance, and Jonathan Guryan for his insightful comments during early discussions on this paper. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence

Andre Nickow, Philip Oreopoulos, and Vincent Quan

NBER Working Paper No. 27476
July 2020
JEL No. I2,J24

ABSTRACT

Tutoring—defined here as one-on-one or small-group instructional programming by teachers, paraprofessionals, volunteers, or parents—is one of the most versatile and potentially transformative educational tools in use today. Within the past decade, dozens of preK-12 tutoring experiments have been conducted, varying widely in their approach, context, and cost. Our study represents the first systematic review and meta-analysis of these and earlier studies. We develop a framework for considering different types of programs to not only examine overall effects, but also explore how these effects vary by program characteristics and intervention context. We find that tutoring programs yield consistent and substantial positive impacts on learning outcomes, with an overall pooled effect size estimate of 0.37 SD. Effects are stronger, on average, for teacher and paraprofessional tutoring programs than for nonprofessional and parent tutoring. Effects also tend to be strongest among the earlier grades. While overall effects for reading and math interventions are similar, reading tutoring tends to yield higher effect sizes in earlier grades, while math tutoring tends to yield higher effect sizes in later grades. Tutoring programs conducted during school tend to have larger impacts than those conducted after school.

Andre Nickow
Northwestern University
Global Poverty Research Lab
601 University Place
Chicago, IL 60208
a-nickow@kellogg.northwestern.edu

Vincent Quan
Abdul Latif Jameel Poverty Action Lab
North America (J-PAL North America)
400 Main Street, E19-201
Cambridge, MA 02142
quanv@mit.edu

Philip Oreopoulos
Department of Economics
University of Toronto
150 St. George Street
Toronto, ON M5S 3G7
CANADA
and NBER
philip.oreopoulos@utoronto.ca

Introduction

PreK-12 tutoring interventions—defined here as one-on-one or small-group instructional programs—rank among the most widespread, versatile, and potentially transformative instruments in today's educational toolkit. As school systems across the globe expand and engage with increasingly diverse student populations, the importance of tutoring continues to grow. Researchers have accumulated a wealth of rigorous evidence on the causal effects of tutoring interventions over the past four decades. Given the widespread use of these programs and the robust body of empirical evidence on tutoring, we believe that synthesizing causal evidence on the effects of tutoring constitutes a key priority for education researchers and practitioners. In the present article, we contribute to this endeavor by developing a unified analytical framework of tutoring programs and using it to guide a meta-analysis of the randomized controlled trial (RCT) evidence. Ours is the first meta-analysis of experimental findings to approach and encompass tutoring interventions as an integrated field of theory and practice.

While researchers and educators have engaged widely with tutoring programs since long before the advent of contemporary education systems, the 1980s saw the emergence of a distinct body of evaluation research focused on these interventions. Most famously, Benjamin S. Bloom (1984) highlighted the challenges and opportunities of tutoring interventions through his clarion call for educators and education researchers to address the "2 sigma problem." Bloom presented evidence from small-scale randomized experiments conducted by two doctoral students (Anania 1983; Burke 1983) demonstrating that one-on-one instruction can generate learning gains of roughly two standard deviations (SD) relative to conventional classroom instructions. The "problem" referred to in Bloom's title is that individual instruction is typically much costlier than group instruction. In the ensuing decades, efforts have intensified to develop and test tutoring

interventions capable of improving learning outcomes within the stark budgetary constraints of real-world education systems.

Since Bloom's exposition, a rich body of empirical work on tutoring interventions has emerged that consistently highlights these programs' potential to yield strong impacts. For example, a recent meta-analysis of different educational interventions targeting elementary and middle school students of low socioeconomic status (SES) found that tutoring was associated with an effect size of 0.36 SD on learning outcomes, the largest effect size of all of the 14 intervention types included (Dietrichson et al., 2017). Yet tutoring program models and their costs vary widely, and education scholars have thus far lacked a framework for systematically comparing the full range of preprimary through secondary tutoring programs relative to one another. Such a framework and associated empirical synthesis could enable the development of a nuanced and empirically grounded understanding of the conditions under which alternative tutoring models may be most effective for different types of students. Mobilizing empirical insights could in turn enable substantial efficiency gains by guiding policymakers and practitioners as they strive to select the most effective options from a range of potential alternatives.

In this article, we construct such a framework to address two research questions:

1) What are the impacts of preK-12 tutoring interventions on learning outcomes?
2) How do effects vary by program characteristics and intervention context?

The study represents the first systematic review and meta-analysis to encompass experimental research on preK-12 tutoring interventions of all types on which experimental studies have been conducted. The last meta-analysis of tutoring interventions was published more than a decade ago and focused exclusively on RCT studies of volunteer tutoring programs that were not geared toward English Language Learner (ELL) students (Ritter et al. 2009). In contrast, one of

the central goals of the present article is to improve knowledge of the relative effectiveness of different tutoring models operating in a wide variety of contexts. Within the past decade, dozens of new studies on tutoring interventions have been conducted. Ritter et al.'s sample included 21 studies, whereas ours includes 96. Moreover, the rigor and sophistication of meta-analytic methods have increased substantially (Cooper et al. 2019). Given the proliferation of meta-analyses within education impact evaluation research, it is surprising that the gap in meta-analyses of tutoring has remained for so long.

We find that tutoring programs yield substantial positive impacts on learning outcomes, with an overall pooled effect size estimate of 0.37 SD. While impacts are significant across most tutoring characteristics, effects are stronger on average for teacher and paraprofessional tutoring programs than for nonprofessional and parent tutoring. Effects also tend to be strongest among the earlier grades. While overall effects for reading and math interventions are similar, reading tutoring tends to yield higher effect sizes in earlier grades while math tutoring tends to yield higher effect sizes in later grades. Tutoring programs conducted during school tend to have larger impacts than those conducted after school. Studies with weaker effects tend to arise from programs like parent and after-school tutoring programs in which it is more difficult to ensure that the tutoring actually occurs. If treatment on the treated estimates were possible to calculate in such cases, effects would likely come up as substantially higher.

In the next section, we elaborate the conceptual framework that we use within the context of the contemporary tutoring policy environment. For illustrative purposes, our theoretical discussion is interspersed with examples from prominent tutoring programs that have been rigorously evaluated. We then explain our methodological approach to analyzing the literature in the third section, before presenting empirical results in the fourth. The fifth section concludes by

discussing a subsample of recent, large-scale impact evaluations, contextualizing our findings against the backdrop of effect sizes for comparable programs, and outlining policy lessons and areas for future research.

Conceptual framework

We conceptualize tutoring as a form of education technology for improving learning efficiency. Education policymakers and practitioners seek to optimize the use of any given education technology by maximizing learning outcomes net of costs. Randomized evaluations generate statistically unbiased estimates of the learning impacts of interventions using the technology in question. Policy-oriented social scientists iteratively develop actionable theories about why some tutoring interventions have more impact or are more cost-effective than others, and how effects vary across contexts. Meta-analysis allows for more formal quantitative testing of effect size distributions, as captured by pre-existing studies, that can help to structure and increase the rigor of theoretical and policy inferences drawn from literature reviews.

In the present section, we describe the conceptual framework from which we approached our analysis. We hope this framework will contribute toward the development of a body of actionable tutoring theories that identify the combinations of tutoring intervention characteristics that are best suited to different educational and socioeconomic contexts. The defining feature of tutoring interventions within our conceptualization is the implementation of one-on-one or small-group academic instruction aimed at supplementing, rather than replacing, classroom-based education. For the purposes of the present analysis, the primary goal of tutoring is to improve the efficiency and equity of student learning outcomes. Like most technologies, use of tutoring

interventions may entail costs of at least three types: investment costs, opportunity costs, and negative externalities.

In designing tutoring interventions and deciding where to scale up, practitioners confront the optimization problem of maximizing learning outcomes with respect to costs and negative externalities. Figure 1 depicts some of the mechanisms by which the learning production function associated with tutoring operate and how characteristics of a tutoring program could shape the distribution of effect sizes.

[Figure 1 about here]

Mechanisms of impact

The tutoring interventions examined in our review attempt to improve student learning outcomes by supplementing classroom-based education. In particular, the majority of interventions cater to students who perform below particular thresholds. Why might tutoring interventions be expected to improve learning in this context? One possibility is that tutoring helps students who have fallen behind by simply providing them with more instruction time. Thus, additional instruction time constitutes one mechanism through which tutoring may improve learning outcomes. On the other hand, if students are pulled out of reading classes for literacy tutoring or math classes for math tutoring, the implicit assumption is that tutoring sessions generate improved learning outcomes for target students in a given unit of time than classroom education. This belief may in turn arise from several theoretical propositions.

Perhaps the most prominently considered mechanism within the literature, aside from additional instruction time, is the *customization* of learning. An already-robust and still-growing

literature has established the pivotal importance of "teaching at the right level" in shaping education outcomes (Banerjee et al., 2015). When students in a classroom span a wide range of skill levels, teachers struggle to address the needs of all at once. Students who miss out on foundational knowledge and skills tend to fall farther and farther behind, and are less able to follow along in class. The productivity of classroom time may thus decline as skill level variation increases. Within this context, learning productivity will increase to the extent that instructional content matches the skill deficits binding the students' learning (Ander et al., 2016), a situation that can be remedied by decreasing students' skill variation though tracking systems or reducing class sizes. Tutoring interventions can be seen as an extreme case of class size reduction in which the class size is reduced to one or a few students. This reduction leads to a massive increase in customization—albeit usually for only a few hours each week—as a supplement to the lower-customization classroom setting.

Beyond customizing learning content, tutoring interventions may also embody a pedagogical moment that is fundamentally distinct from classroom education. One-on-one and small group settings may, for instance, allow for more engagement and rapid feedback, enabling educational activities that would not be possible in the classroom. Without being lost in the crowd of the larger classes, students may approach time spent in tutoring interventions with a greater degree of focus and effort than in classrooms. And, because there are presumably fewer distractions during tutoring sessions than regular classes, students may spend a larger share of time on task in tutoring sessions than in classes.

Another potentially important element of tutoring interventions is the human connection generated by tutor-student relationships, i.e., the mentorship relationship. Mentorship programs represent a distinct (although partially overlapping) field of practice from tutoring interventions,

but tutoring programs may engender mentorship relationships that go beyond the academic content of the tutoring session. While it could be that the mentorship relationship represents an effectively separate dimension from the tutoring relationship, there may be interaction effects by which positivity associated with the personal mentorship relationship may carry over to positivity toward the educational content, or toward the academic learning process more broadly.

Additionally, tutoring programs may yield positive externalities for students who do not receive the tutoring. This yield could occur, for instance, to the extent that tutoring programs decrease the sizes of the classes in which the tutored student would otherwise be tutored, and increase peer learning effects to the extent that tutored students' outcomes improve. Another potential mechanism is that tutoring programs, particularly when implemented at earlier grade levels, may separate students who had fallen behind as a result of circumstance from those with specific learning disorders (Schwartz, 2005, p. 257). While externalities of this sort have rarely been experimentally tested in practice, they constitute a potentially important dimension to tutoring impacts at scale to keep in mind when considering policy inferences.

Characteristics of tutoring programs

Tutor skills and qualifications

Given the above set of potential mechanisms, what elements of tutoring programs are most likely to shape impacts? The most prominent input for tutoring interventions is typically the tutor's human capital or skill, i.e., the extent to which the tutor's behavior leads to higher learning gains, holding other elements of the intervention constant. We expect that more highly educated, trained, and experienced tutors will have stronger tutoring skills and will demand higher wage premiums in exchange for the associated impact premium. In other words, interventions employing more

highly-qualified tutors will likely be more expensive, but also more effective. We thus expect *tutor type* to moderate the impacts of tutoring interventions.

Four broad categories of tutor type emerged inductively from our review of the literature: teachers, paraprofessionals, nonprofessionals, and parents. In *teacher tutoring* interventions, certified classroom teachers fulfill the role of the tutor. The most prominent teacher tutoring program in the literature is Reading Recovery[1], developed by University of Auckland educational psychologist Marie Clay and piloted it in New Zealand in 1979. The program launched in the United States during the 1984-5 school year in the Columbus Public Schools, and in 1985-6 expanded to 12 public schools, before spreading widely across the United States and beyond. Training for Reading Recovery tutors is intensive. Tutor candidates—who are already trained and certified teachers—must undergo a graduate-level course lasting a full year, followed by continuing consultations and other professional development activities (Sirinides et al., 2018).

Moving to the next category, *paraprofessional tutoring* interventions employ tutors who are professionally engaged in their tutoring roles but who are not certified teachers. There have been a much wider array of paraprofessional tutoring programs implemented and tested than teacher tutoring. This category of tutors includes school staff members, undergraduate and graduate students in the education field, and fellows in professional development and service programs. One prominent paraprofessional tutoring program within the literature is Number Rockets[2] (known in some versions as Galaxy Math), a first-grade math program. When implemented at smaller scales as part of exploratory research studies, tutors were mostly graduate student research assistants (Fuchs et al., 2005; 2013), whereas school staff members were employed as tutors in a large-scale impact evaluation of the program (Gersten et al., 2015).

[1] https://readingrecovery.org/
[2] https://frg.vkcsites.org/what-are-interventions/math_intervention_manuals/

AmeriCorps[3], a US government-funded service-oriented fellowship program aimed at recent high school or college graduates, represents another important source of paraprofessional tutors. AmeriCorps fellows have served as tutors in programs ranging from the early elementary Minnesota Reading Corps (Markovitz et al., 2014) and late elementary Minnesota AmeriCorps math tutoring (Parker et al., 2019) to Saga Education early secondary math tutoring, which now operates in Chicago, New York City, and Washington, DC.[4] Fellows remain in programs of this type for typically only a year or two (AmeriCorps lasts for ten months), but receive between several days' and a few weeks' worth of training as well as close supervision.

Nonprofessional tutoring interventions deploy volunteers who are not professionally engaged within the education field, including community residents and retired adults. These interventions are often referred to as *volunteer tutoring*, but we use the word nonprofessional to distinguish these interventions from paraprofessional tutoring and because receipt of compensation is not the defining feature of nonprofessional tutoring interventions within our framework. Reading Partners[5] is one nonprofessional tutoring program that operates in ten states across the US, drawing on the services of AmeriCorps fellows (in addition to permanent staff members) to hire and supervise tutors, but uses unpaid community volunteers who receive only about an hour of training to do the actual tutoring. Meanwhile, Experience Corps[6], a nonprofessional tutoring program implemented by the AARP Foundation, matches schools with adults over the age of 50 who tutor children in early elementary grades on reading.

Parent tutoring interventions provide instruction and guidance to parents or other guardians to tutor their children at home, outside of school hours. These interventions have not

[3] https://www.nationalservice.gov/programs/americorps
[4] https://www.sagaeducation.org/our-story
[5] https://readingpartners.org/
[6] https://www.aarp.org/experience-corps/

been experimentally studied or implemented as frequently or at as large a scale as have programs from the other categories, perhaps in part because parent tutoring may be subsumed within broader parental involvement programs. Parent tutoring programs typically provide parents with training and/or instructional materials that recommend a specific tutoring approach and ask them to commit to tutoring at regular intervals. One tutoring strategy commonly employed in parent as well as nonprofessional tutoring programs, given the limited training required, is "paired reading" (Topping, 1986), which involves alternating between the tutor and tutee reading together and the tutee reading alone as the tutor listens for errors. For instance, a recent program in Hong Kong provided parents with 12 sessions on tutoring using paired reading over seven weeks, during which times the parents were asked to practice the tutoring method at least four times (Lam et al., 2013). A paired reading program in Switzerland provided parents with two training sessions of about 90 minutes each, and asked them to practice paired reading tutoring with their children two to three times per week, for approximately 20 weeks, at home outside of school hours (Villiger et al., 2019).

While real-life tutoring programs involve a diverse array of individuals who may span more than one of these categories, tutoring interventions implemented at scale generally specify particular categories of individuals when planning for recruitment. We expected *a priori* that teacher tutors would be the most highly skilled, followed by paraprofessional tutors, followed by nonprofessional and parent tutors. Teachers and paraprofessional tutors have been explicitly trained in education, while nonprofessional and parent tutors generally have not. Teacher tutors in the literature we reviewed typically had substantially more training and experience than did paraprofessional tutors. While specialized practitioners with higher-level graduate degrees may have acquired stronger tutoring skills even than classroom teachers, no programs included in our study were designed for tutors at that professional level.

Teacher tutoring interventions likely incur the highest salary or wage costs of the four categories that we included. Certified teachers could presumably be instructing entire classrooms and receiving the accompanying salary in the time that they spend teaching individual students or small groups as tutors. Paraprofessionals are sometimes paid salaries or stipends and may incur other costs, but on average require significantly less compensation than teachers. Nonprofessional tutors are usually unpaid or, if compensated at all, receive small honoraria or travel reimbursements. No studies we examined offered compensation to parents for tutoring. Where teacher or paraprofessional tutors are drawn from within schools, they may incur relatively low administrative costs since they are already embedded within the school's organizational structure. Paraprofessional and nonprofessional tutoring programs involving external tutors may incur higher administrative costs. Administrative costs for parent tutoring interventions may typically be low since in general there is less scope for supervision of parents than for other tutor types.

One family of interventions that we elected not to include within the meta-analysis is that of *peer* or *cross-age* tutoring. Like the tutoring interventions discussed in the present paper, these interventions consist of one-on-one or small-group instruction. However, unlike the interventions we include, the "tutors" for peer and cross-age interventions are classmates or schoolmates of the tutees. By far the most prominent and rigorously tested intervention model in this category within the contemporary literature has been Peer-Assisted Learning Strategies (PALS). In the PALS model, classes are typically broken into student pairs with one student temporarily assigned to be the tutor and the other to be the tutee (Fuchs et al., 1997).

While the spirit of tutoring is clearly manifest in these contexts, from the perspective of this review we see PALS to be more of a collaborative learning intervention than a tutoring intervention as described above, because it is designed with both the tutor's and tutee's learning

in mind. Cross-age tutoring interventions, in which tutors are in a higher grade than tutees, are perhaps more similar to conventional tutoring interventions of the type we study, but even these typically involve a pedagogical approach aimed at benefiting the tutor as well as the tutee. We briefly discuss findings from the literature on peer and cross-age tutoring in the "Contextualizing PreK-12 Tutoring Programs" section below.

Curriculum characteristics

Besides tutor skills, the effectiveness of tutoring programs may depend substantially on the content that is being taught. First, tutoring interventions may cover different subjects. The programs studied in our sample all fell into the categories of literacy (i.e,. related to language or reading) or math. Secondly, curriculum within these subjects also changes enormously across grade level. When tutor skill, program and contextual characteristics are held fixed, effect sizes may differ depending on the subject of the tutoring intervention and the grade level targeted.

Even within particular age-subject curricula, teaching strategies and content may vary across tutoring programs. For instance, some early reading programs focus more on phonics, while others focus more on comprehension. Additionally, some tutoring programs provide higher levels of structure and more detailed guidance for tutors than other programs. At one end of the spectrum, programs like Reading Recovery and Number Rockets consist of highly structured lessons with clear and detailed directives to the tutors. In Reading Recovery, each lesson begins with "rereading familiar books" aloud, followed by targeted letter and word recognition activities, story composition, and reading a new book, which students are then "expected" to practice reading at home (Sirinides et al., 2018, p. 317). While curriculum and pedagogical approach are tightly controlled, teachers are expected to draw upon their extensive specialized training to customize

lessons to a student's particular strengths and weaknesses. Meanwhile, Number Rockets consists of about 45 scripted lessons designed for delivery over the course of about 17 weeks, and provides comprehensive instructions for tutors (Gersten et al., 2015). On the other end of the spectrum, programs like the Northern Ireland nonprofessional elementary reading program Time to Read provide minimal structure and significant leeway to tutors (Miller & Connolly, 2013).

While subject and grade level are included within our quantitative meta-analyses, pedagogical approach and level of structure were not possible to reliably code across the full range of studies that we collected. We consider these factors qualitatively in interpreting the data.

Mode of delivery

Another potentially important set of influences on the success of tutoring programs are those associated with delivery mode. We highlight two variables falling within this category: first, tutor-student ratios, and then timing and location of the tutoring.

Tutoring may be conducted one-on-one where a tutor instructs an individual student, or tutors may instruct pairs or small groups of students simultaneously. All else being equal, interventions with more students instructed by a single tutor at the same time should incur lower costs. More students per group may reduce tutoring impacts if it means dividing the tutors' time and attention. On the other hand, pairs or larger groups may improve tutoring programs to the extent that there are gains to group learning in the program areas, or if tutoring in pairs or small groups reduces a sense of stigma.

Nonprofessional tutoring programs are typically one-on-one, which could in part arise from the fact that greater skill may be required to tutor more than one student at a time. Parent tutoring programs are typically designed as one-on-one, since parents usually have only a single child at a

particular grade level. Intuitively, it would seem that teacher tutoring would fit well with small groups as well as one-on-one tutoring, given teachers' experience with teaching multiple students at once. In practice, however, Reading Recovery is the only teacher tutoring program that has been widely implemented and evaluated, and Reading Recovery calls for one-on-one tutoring. Most of the variation in one-on-one vs. small-group tutoring thus occurs within paraprofessional tutoring. Of the paraprofessional tutoring program models discussed thus far, three (Minnesota Reading Corps, Minnesota AmeriCorps, and Saga Education) tutor pairs of students, while one (Number Rockets) accommodates groups of two to four students per tutor in each session.

In addition to the number of students participating in each session, when and where tutoring occurs may substantially shape its effects. One key distinction here is whether tutoring is held during or outside of school hours. This question is in part a function of tutor type. Almost all parent tutoring occurs at home, outside of regular school hours. On the other hand, while after-school teacher tutoring is certainly conceivable if held at school, Reading Recovery is conducted at school during school hours, and we have not come across rigorous evaluations of any after-school teacher tutoring programs. In our review, variation in the time and location of tutoring occurs within paraprofessional and nonprofessional tutoring programs. Of the paraprofessional and nonprofessional tutoring programs discussed so far, Number Rockets, Saga Education, Time to Read, and the Minnesota reading and math programs all take place during school hours. Reading Partners may take place either during or after school.

While parent tutoring usually occurs at home, paraprofessional and nonprofessional tutoring programs that occur outside of school hours are typically held at school. For instance, the Swiss paired reading study discussed above with reference to parent tutoring also included a nonprofessional "volunteer" experimental arm. While tutoring sessions for both groups occurred

after school, parent tutoring occurred at home and volunteer tutoring occurred at school. As the authors point out, these differing locations may yield substantially different tutoring environments; for instance, the home environment may be more comfortable and relaxing, but also more distracting and less structured (Villiger et al., 2019, p. 56). Mattera et al. (2018) evaluate a small-group kindergarten math tutoring program called High 5s, which is introduced to students as a math club and occurs at school outside of regular class hours. Sometimes after-school tutoring programs are held at community centers (Morris et al., 1990), or in students' homes, as with a California foster tutoring program (Zinn & Courtney, 2014).

Important timing distinctions may exist even among programs held during school hours, particularly whether tutoring sessions replace classes of the same topic, classes of different topics, or recreational (or otherwise unfilled) time. Ideally, tutoring sessions would replace whichever time slots exhibit the lowest opportunity costs, but this timing may be difficult to discern and coordinate for individual students, let alone across classes and schools. Both Reading Recovery and Number Rockets ask that schools schedule tutoring sessions to avoid schedule conflicts with the respective subject of tutoring. Thus, treated students in Reading Recovery are pulled out of other classes, such as math, or recreational activities to attend reading tutoring, and vice versa for Number Rockets. In other cases, students are pulled out of the class associated with the subject of tutoring (i.e., reading classes for reading tutoring and math classes for math tutoring). Control group students generally remain in regular classes and are not pulled out. For after-school programs, treated students are presumably missing time for homework or extracurricular activities that control students engage in. Studies rarely report in detail what specific activities treated students would have engaged in had they not been tutored at that time, or which supplementary services struggling control group students may have received, highlighting areas to which future

evaluations should pay close attention. When considering the program impacts presented below, it is important to consider the counterfactual in terms of the educational or recreational time used up by tutoring. This holds for after-school as well as during-school interventions, given that the former may also involve tradeoffs in terms of homework time, recreation, or other developmentally beneficial activities.

Dosage

Tutoring programs vary widely in terms of frequency and length, as well as program duration and overall number of lessons. Program models generally call for tutoring between one and five days per week. Sessions vary in their length from 10-15 minutes to more than an hour, with most programs suggesting sessions of between 30 and 60 minutes. More days per week may be expected to increase effect sizes unless and until the tutoring sessions begin to crowd out other learning inputs beyond a particular threshold. Similarly, net of opportunity costs, one might expect longer sessions to yield higher effect sizes until the point at which students' attention span becomes an issue. Reading Recovery calls for daily 30-minute sessions, while Number Rockets calls for 40-minute sessions three or more times per week. At the high end, Saga Education tutoring consists of daily 60-minute tutoring sessions.

Overall program durations may vary in length from several weeks to one or two school years, although the majority of the prominent tutoring programs that we reviewed lasted between ten weeks and one school year. Some interventions are designed such that students who improve more quickly in outcomes associated with the tutoring subject area are released from the program earlier. For instance, Reading Recovery may last between 12 and 20 weeks, depending on the

speed of the student's demonstrated reading improvement. All else being equal, longer intervention periods should yield higher effect sizes.

Meta-analysis methodology

To synthesize findings from the experimental research on tutoring, we draw on methodological best practices for meta-analyses that have coalesced over the past several years (Pigott & Polanin 2019; Siddaway et al. 2018). We outline our methods and analytical approach in this section.

Study and estimate inclusion criteria

Throughout the majority of this paper, we take the "study" as the main unit of analysis. We define a study as the enactment of a research design with a particular sample over a prospectively planned time horizon. In some cases, results from a single study may be reported in multiple articles, e.g., when preliminary papers report on the first cohort of a study while implementation for subsequent cohorts is in progress, or when dissertations or working papers are revised and published in peer-reviewed journals. In implementing our systematic review, we collected all eligible papers and aggregated them by study to avoid double counting.

The vast majority of experimental tutoring studies report multiple estimates (i.e., impact of the program on more than one outcome measures and/or the impact of more than one treatment arm). We identified several criteria to decide which estimates to include in our meta-analysis. A study was included if it contained one or more eligible estimates. Our criteria were as follows. First, we included only estimates evaluating treatments that met our definition of tutoring interventions as compared to a non-tutored control group. As mentioned above, we define tutoring for the purposes of this paper as one-on-one or small-group human (i.e., non-computer) instruction

aimed at supplementing classroom-based education. We excluded studies that lacked one or more treatment arms in which the tutoring invention was the only treatment, i.e., studies that identify only the effects of tutoring when bundled with other intervention components. For example, we excluded studies in which the only tutored treatment arms also included computer-based activities, or other non-tutoring remedial activities. We also excluded studies that exclusively tested alternative tutoring interventions against one another, i.e., studies lacking a non-tutored control group. Studies were also excluded if they consisted of fully individualized instruction of the topic in question (e.g., Anania 1983; Burke 1983) since we conceptualize tutoring interventions for the purposes of this paper as a class of technology used for supplementing, rather than replacing, school-based learning. This boundary separates tutoring interventions as conceived in this article from homeschooling, for instance.

Second, we included only studies at the preschool through secondary level. While programs otherwise meeting our definition of tutoring interventions have been implemented and tested in the context of higher education and professional certification programs, post-secondary programs have tended to focus on specialized skills rather than broader academic learning outcomes. Furthermore, tutoring interventions in primary and secondary contexts are likely to remain the key goal for researchers and practitioners focused on improving equity within educational systems.

Third, we included only studies presenting impact estimates based on randomized controlled trials (RCTs). While observational and quasi-experimental studies have contributed a great deal to debates surrounding tutoring, and figure prominently within our broader narrative discussions, including non-experimental studies within our formal meta-analysis would have introduced the need for assumptions surrounding potential bias. While debates surrounding these

assumptions are worthwhile in some contexts, more confidence can be placed in the internal validity of experimental studies. The RCT evidence on tutoring has grown more than enough to justify an independent examination of the findings, and then interpret results from them in the context of a wider research and policy environment.

Fourth, we included only studies presenting impact estimates on learning outcomes. We excluded studies focusing exclusively on behavioural outcomes such as attention or disruptive behavior. Fifth, we included only studies that have been published since 1980. While a handful of relevant studies were conducted before 1980, the vast majority have been conducted since then. Educational systems and experimental research standards have changed substantially over the past few decades. Sixth, we included only studies that presented the data required to compute effect sizes (the calculation procedure is described below).

Search protocol

In order to identify the full population of eligible studies, we searched several types of databases. First, we searched academic databases containing peer-reviewed research studies and scholarly working papers (EBSCO, J-PAL, JSTOR, NBER, SCOPUS, SSRN, and Web of Science). Second, we searched the primary database compiling university theses and dissertations (Proquest Dissertations). Finally, we searched databases and organizational websites containing professional development reports and other forms of gray literature (American Institutes for Research, Cochrane Library, IPA, Mathematica, MDRC, and NORC).

We searched each of these databases using the search terms “tutor* & random*” (with asterisks indicating wildcard) within each of these databases, or the closest equivalent given the specific setup of each database or website. This procedure constituted our primary search tool. To

catch any articles that may have been missed, we conducted backward and forward bibliographic searching for each article included following the above search procedures, looking through past articles cited in each included study as well as future articles that cited each included study. The main searches were conducted in September-October of 2019, and additional searches were conducted in February 2020 to identify documents that had been made public since October 2019.

Once each search had yielded its results, we screened the studies identified in two stages. During the first stage, we examined titles and abstracts and removed articles that met any of the exclusion criteria described above. All articles that passed this first stage were read and considered within the broader article and narrative review, but were excluded from the meta-analysis. We then subjected each article to a second stage of review in which we read the full text and checked whether each of the inclusion criteria had been met. Coding was conducted by one of the authors and checked by research staff members. Each area of disagreement was discussed and resolved.

Calculating effect sizes

The studies included in our review present a wide variety of estimates. Following recent consensus on meta-analysis best practices (Pigott & Polanin, 2019, p. 8), we calculate effect sizes for inclusion in the meta-analysis using treatment and control group means, standard deviations, and sample sizes. Where available, we use adjusted means along with unadjusted standard deviations following Dietrichson et al. (2017, p. 255). We noted unadjusted means where adjusted means were unavailable, and we used author-reported estimates where even unadjusted means were not presented in the articles. Because the outcomes of interest in the studies reviewed here are continuous and rely on a variety of different outcome scales, we calculate standardized mean

difference effect sizes. We use Hedge's *g*, which approaches Cohen's *d* for larger sample sizes and corrects for bias in smaller samples (Borenstein 2009).

We include estimates of only academic learning outcomes, thus excluding perceptions, behavior, attention, and other outcome categories. In a few of the included studies, learning outcomes in subject areas that are unrelated to the tutoring intervention were included to test for negative spillover effects into other subjects—these are excluded from our analysis. Since the vast majority of studies report only post-test outcomes immediately following treatment, we code whether each outcome was measured three months or less following treatment and include only those in that time frame in our main analyses.

Meta-analysis and meta-regression models

We analyze a series of meta-regressions and supplemental meta-analyses as the study's primary sources of inference. These regressions are descriptive in the sense that we are concerned with identifying associations between potential moderators of tutoring and learning effect sizes, rather than making strong causal statements. Furthermore, we describe the population of existing studies of tutoring interventions rather than a probability sample of all tutoring programs. Nonetheless, the meta-analyses provide us with key quantitative benchmarks for identifying the magnitude and significance of the relationships hypothesized within our framework. Our regression models follow the approach outlined by Tanner-Smith & Tipson (2014), and utilizes their Stata program *robumeta*. Specifically, we deploy random effects models with inverse variance weights. The weights increase the influence of studies with larger sample sizes and greater precision. We account for independence of effect sizes within studies using robust variance estimation (Hedges et al., 2010).

We begin with single-variable regressions for all estimates, and then for subsets of estimates delimited by study and treatment characteristics. Tutoring intervention characteristics that can be conceptually distinguished may still tend to cluster together in practice, and these analyses allow us to observe pooled effect sizes and their variation across characteristics. We then add in sets of potential moderators as control variables and observe their associations with effect sizes in the context of multivariate meta-regression to find suggestive evidence for the finer-grained dynamics that attempting to pull apart these variables may reveal. Analysis of the quantitative results is accompanied by narrative analysis of the literature that fills in gaps on subtle characteristics that could not be reliably quantified and contextualizes clusters of results within research and policy contexts.

Findings

Descriptive analysis

[TABLE 1 ABOUT HERE]

The search and screening process described above yielded a total of 96 studies. Table 1 shows the breakdown of the studies over a variety of intervention and study characteristics, disaggregated by subject and tutor type. Each cell shows the frequency of studies falling into the categories defined by its respective row and column, as well as its proportion vis-à-vis the full sample of studies included within the meta-analysis.

As shown in Table 1, literacy tutoring is far more common within our sample than math tutoring, with nearly 80% of studies evaluating a tutoring intervention with a literacy component,

and just over a quarter evaluating math tutoring interventions. Paraprofessional tutoring accounts for the largest share of tutor type at nearly half, followed by nonprofessional, teaching, and then parent tutoring. Almost all math interventions utilized paraprofessional tutors. Tutoring interventions in our sample cluster overwhelmingly within (pre-middle) elementary school, with only 7% of interventions involving students in sixth grade and above. Almost half of all studies involve first-graders, with this disproportionate concentration largely resulting from reading interventions since students typically learn to read in first grade.

For delivery mode, tutoring interventions administered during school are far more common in our sample than after-school programs, represented in roughly 80% and 20% of studies respectively. Most of the variation here comes from paraprofessional and nonprofessional tutoring, since all teacher tutoring interventions in our sample occurred during school and all but one parent tutoring intervention occurred outside of school hours. Roughly 70% of studies include one-to-one tutoring, while about a quarter of them include one or more treatment arms with three or more students per tutor.

A relatively small but non-negligible handful of studies look specifically at effects of tutoring interventions for English Language Learners (ELL) and foster students. The literature on tutoring for foster children has been relatively self-contained, whereas ELL students are present in varying concentration in many of the studies, and even studies focusing explicitly on ELL learning fit within the broader non-ELL focused literature. We coded studies into the ELL category if they specifically discussed ELL learning and/or if the half of the sample or more were identified as ELL students.

Finally, the vast majority of studies were (eventually) published in academic journals. This category is especially large because we coded dissertations, evaluation reports, working papers, and other write-ups later published as academic journal articles into the latter category.

Study characteristics and pooled effects

[FIGURE 2 ABOUT HERE]

[TABLE 2 ABOUT HERE]

We next turn to the central task of our quantitative meta-analysis: estimation of pooled effect sizes. Figure 2 depicts a forest plot with effect sizes for each study, averaged by treatment arm. Table 2 presents pooled effect sizes, standard errors, and sample information for all studies as well as for sub-samples of studies defined by sample size, publication year, publication type, and risk of bias.

The estimates shown in the first row of this table represent the primary answer to this study's first research question, i.e., what is the causal effect of tutoring interventions on learning outcomes based on findings from experimental studies? We find that, across all estimates and studies included in our analysis, tutoring interventions show a statistically significant and substantively large effect size on learning outcomes of 0.37 SD. Remarkably, this estimate is almost identical to the pooled effect size of 0.36 SD found by Dietrichson et al. (2017) in their meta-analysis on education interventions for students of low socioeconomic status (SES), and similar to the 0.30 SD effect size found by Ritter et al. (2009) in their review of experimental studies on volunteer tutoring. The similarity is especially striking given that their sample of

tutoring studies included only tutoring interventions for low-SES students whereas ours included all K-12 tutoring interventions that otherwise met our criteria, and we included only randomized experiments while their inclusion criteria allowed studies of any type of treatment-control group design. Our pooled impact estimate is also nearly identical to the effect sizes found in multiple evaluations for the math tutoring program Number Rockets (Fuchs et al. 2005; 2013; Gersten et al. 2015), which was not included in Dietrichson et al.'s (2017) review.

The sample size panel within Table 2 shows that effect sizes remain broadly consistent for studies with sample sizes of up to around 400. The pooled effect size for studies with samples greater than 400 is roughly a quarter of a standard deviation. However, a close look at effect sizes for larger-sample studies shows that effect sizes do not continue to fall with sample size, but instead plateau and remain consistent after the 400 mark. Large effect sizes within small-sample studies are explained mostly by literacy tutoring outliers, while math program effects remain more consistent.

Findings are generally consistent across publication years, with the most recent decade showing a slight decline. Effect sizes were similar between those studies that have been published in academic journals and those that have not. The fact that unpublished papers have larger effect sizes than published articles is encouraging in that this pattern is in the opposite direction as might be expected if publication bias were a substantial issue.

In general, the studies included in this review were of consistently strong quality. Limiting the sample to RCTs went a long way toward eliminating many of the issues pertaining to bias and variation in study quality that many analyses face. Nonetheless, we follow best practice recommendations in coding studies for "risk of bias" (Pigott & Polanin, 2019, p. 7). Our criteria are broadly inspired by criteria at the heart of the Cochrane risk of bias framework (Conn, 2017,

p. 869; Higgins et al. 2011), but we focus specifically on three dimensions for which we were able to detect some degree of variability, and that we expect to be correlated with risk of bias given association with overall study quality: the extent to which studies systematically reported information on 1) the intervention, 2) study design, and 3) relevant statistics.

We created a bias risk index ranking studies from one to three on each of the above-mentioned three dimensions. Summing these scores yielded a nine-point index, with one representing the highest risk of bias and nine representing the lowest. While we feel that all studies in our sample exhibit a low risk of bias, we constructed a dummy variable to identify those studies with scores of 8 or 9 on the quality index. The bias risk index panel of Table 2 shows that effect sizes are similar regardless of bias risk, but that studies with higher bias risk have slightly higher effect sizes overall. Still, studies coded as having "low risk of bias" show a pooled effect size that is nearly identical to the overall pooled effect size from the top left corner, showing that study quality and bias risk do not affect our overall estimates.

Program characteristics and pooled effects

This subsection and the next address our second research question, i.e., how treatment effects vary with characteristics of tutoring programs and the contexts in which they operate.

[Tables 3A and 3B about here]

Tables 3A and 3B present pooled effect size estimates using the same single-variable meta-regression models as in Table 2. However, these tables categorize studies into subsets defined by characteristics relating to the tutoring interventions, and the context in which they operate, rather than study characteristics as in Table 2. Standard errors or entire estimate cells are omitted where a lack of degrees of freedom precludes reliable use of the robust variance estimation techniques we employ (Hedges et al., 2010; Tanner & Tipton, 2014). Columns in Table 3A separate studies by tutor type and columns in Table 3B separate studies by grade level, with rows in both tables breaking studies into categories based on the remaining characteristics that we coded in our dataset.

The central lesson highlighted by Tables 3A and 3B is that, notwithstanding some variation in magnitude and gaps in the literature, *tutoring interventions exert substantial effects on learning across a wide range of program characteristics*. Effect sizes are positive and educationally significant for the vast majority of subgroups, particularly within teacher and paraprofessional tutoring categories. Lack of significance, when it does occur, is driven more by wide confidence intervals than by small effect estimates. However, the data do show signs of meaningful variation across categories. We begin by describing variation across tutor type and grade, the two variables that structure Tables 3A and 3B, and then describe variation in effect sizes across remaining intervention characteristics in turn.

Tutor type

The first row of Table 3A reveals that *teacher tutoring programs yield the largest impacts, followed by paraprofessional tutoring programs, with nonprofessional and parent tutoring accounting for the lower end of the impact distribution.* However, the advantages of teacher

tutoring over paraprofessional tutoring are driven primarily by first-grade interventions, and the high scores of first-grade teacher tutoring interventions are in turn attributable in large part to studies evaluating Reading Recovery.

Reading Recovery has been subjected to five experimental evaluations meeting our study criteria, spanning three decades and taking place in Ohio (Pinnell et al. 1988; 1994), Australia (Center et al. 1995), and multiple US states (Schwartz, 2005; Sirinides et al., 2018). A recent large-scale evaluation (Sirinides et al., 2018), discussed further below, represents the largest-sample evaluation in our sample. Effect sizes are substantial, ranging from 0.56 in Sirinides et al.'s (2018) large-scale evaluation to 1.09 SD in Center et al.'s (1995) study, the latter among the highest treatment effect size averages in our sample.

Still, effects from other teacher tutoring interventions tend to be high as well. These effects reflect a diversity of training and implementation models, most substantially less training-intensive than Reading Recovery, showing that Reading Recovery is far from the full story of teacher tutoring's success within our sample. Other teacher tutoring programs for early elementary reading also show strong results, ranging from around half of a standard deviation (Blachman et al. 2004; Mathes et al., 2005) to more than a full standard deviation impact from a "multisensory" reading tutoring program in Sweden (Bøg et al., 2019). Two teacher tutoring programs designed specifically for English Language Learners (Borman et al.; Vaughn et al.) show effect sizes of around 0.50 SD on reading in the language of instruction, although nonsignificant results for the other language (the Borman et al. intervention is taught in Spanish while the Vaughn et al. intervention is taught in English). Each of these programs involved daily tutoring, with the one in Borman et al.'s study evaluating an English-Spanish dual-language learner tutoring effort called *Descubriendo la Lectura*, explicitly modeled after Reading Recovery. Evaluations of reading

programs for later grades show substantial promise, but results are more mixed (O'Connor et al. 2002; Vaughn et al. 2019; Wanzek & Roberts).

Only four teacher tutoring studies meeting our criteria focused on math, leaving an unreliable standard error alongside a large coefficient. Smith et al. (2013) test Math Recovery, a program following much of Reading Recovery's structure and framework but adapted for math, with a sample of more than 700 students across two states. They find effect sizes that are slightly smaller than those found for Reading Recovery, at around 0.4 SD. Fuchs et al. (2002B; 2008A) observe some of the largest average effect sizes in our entire sample from a math programs for third- and fourth-graders respectively, but these are small sample studies that focused specifically on story problems. Lorenzo et al. (1993) find no effects, but their study was conducted more than 25 years ago with a small sample that seems to test a relatively unique and informal program, making it difficult to clearly map outcomes onto contemporary debates.

As discussed above, paraprofessional and volunteer tutoring programs each subsume a range of tutor type subcategories. The most common types of paraprofessional tutors in our sample were interventionists employed by the school or community (Clarke et al. 2016A; 2017; Doabler et al. 2017; Gersten et al., 2015; Jenkins et al 2004; Lane et al. 2007; Mattera et al. 2018; O'Connor et al. 2010; Vadasy & Sanders, 2008A; 2008B; 2008C; 2009; 2010; 2011; Vadasy et al. 2006A; 2006B; 2007), undergraduate and graduate students and trainees in education-related fields (Allor & McCathren, 2004; Case et al. 2014; Denton et al., 2004; Fuchs et al., 2005; 2013; Jung 2015; Lane et al. 2009; Mayfield, 2000; Powell et al., 2015; Swanson et al., 2014 Young et al., 2013), participants in postgraduate or civic service programs (Cook et al., 2015; Markovitz 2014), Math (Parker et al., 2019), and research team members employed directly by principal investigators (Fuchs et al. 2019; Gilbert et al., 2013; Toste et al. 2017; 2019; Bryant et al.; Fuchs et al. 2009).

Nonprofessional, or *volunteer*, tutoring programs may employ community volunteers (Al Otaiba et al., 2005; Jacob et al., 2016; Loenen, 1989; Mooney 2003; Morris et al. 1990; Vadasy et al., 1997A; 1997B; 2000), members of business volunteer networks (Baker et al., 2000; Miller et al., 2012; Miller & Connolly, 2013), senior citizens or older adults (Fives et al., 2013; Lee et al., 2011; Rebok et al. 2004; Rimm-Kaufman et al. 1999), undergraduate non-education majors (Lachney, 2002; Lindo et al., 2018; Woo, 2005), and nonprofessionals selected by the research team (Benner, 2003). Given their diversity, paraprofessional and nonprofessional tutoring programs are discussed in the following sub-section separately by grade level and subject.

Finally, parent tutoring interventions consist of providing parents with training, materials, and follow-up support so that they can act as the tutors. While parent tutoring programs may be expected to require less supervision than the other types given the nature of parental autonomy, parents might also have the highest internal motivation in wanting children to succeed. Parent tutoring interventions had the least number of studies devoted to them in our sample, and there were no large-scale impact evaluations of parent tutoring programs. While effect sizes are weaker on average than for teacher and paraprofessional tutoring interventions, even 0.20–0.25 SD can be substantial if costs are low, as they typically are for parent tutoring programs. The largest-sample in this study was reported in Lam et al. (2013), a preschool paired reading tutoring program designed to be administered by parents in Hong Kong involving just under 200 preschoolers. Promisingly, this study showed an effect size close to our meta-analysis's main pooled effect estimate of slightly over a third of a standard deviation.

Panels A–D of Figure 3 depict the studies' unweighted average effect size distributions graphically as kernel density functions for each tutor type. While these plots are purely descriptive, they reveal some additional insights into differences in patterns across tutor types. Most noticeably,

despite the higher average effects of teacher tutoring programs relative to paraprofessional tutoring programs, effect sizes for paraprofessional tutoring exhibit substantially more consistency. This is particularly remarkable given the relatively wide range of individuals who may be classified as "paraprofessional." Several teacher tutoring studies show very large effects, skewing the distribution rightward. The high degrees of variation seen in Panels C and D indicate that these findings may need to be broken up further to identify clear policy projections.

Grade level and subject

The first row of Table 3B reveals that, at least up until middle school, *effect sizes roughly decline with grade level*. There have not been enough studies that meet our criteria to support reliable estimates for parent tutoring, or tutoring in sixth grade through high school across subcategories, or even to calculate a reliable standard error for the sixth grade and up pooled estimate. While PreK-kindergarten interventions tend to have the highest overall effect sizes, the largest differences in pooled effect sizes across other variables tend to be between PreK-kindergarten and first grade versus the rest.

Rows 2 and 3 of Table 3B, however, reveal a striking result: *the pattern of declining returns to tutoring across our grade level categories are explained entirely by literacy programs.* Math programs, if anything, show a reverse trend, with increasing impacts from PreK-kindergarten to first grade to grades 2 through 5.

Overall effect sizes for literacy and math interventions are similar to one another, although comparison is difficult, given a smaller study sample and less diversity for math tutoring studies. Panel D of Figure 3 shows a greater degree of consistency in effect sizes for math programs relative to reading, although this is likely at least in part a function of the smaller number of experimental

studies that have been done on math tutoring. Disaggregating effect sizes by subject in rows 2 and 3 of Table 3A reveals that *the relatively smaller pooled effect sizes for nonprofessional and parent tutoring are driven primarily by reading programs* since the vast majority of math tutoring programs utilize paraprofessional tutors. There are too few math tutoring studies to statistically compare the effects of different tutor types against one another for math. Because paraprofessional and nonprofessional tutoring programs are most spread out across grade levels, and only paraprofessional tutoring has a significant number of math tutoring programs, we focus on interventions in these areas in the remainder of this subsection.

Beginning with reading, Vadasy and collaborators conducted a series of nine randomized studies on elementary school literacy interventions. For these interventions, tutors are hired by school districts from local school communities. Average effect sizes for kindergarten range from just under a half to two-thirds of a standard deviation for kindergarten (Vadasy et al., 2006A; Vadasy & Sanders, 2008A; 2010). Tutoring programs involving explicit instruction (e.g., in phonics, decoding, and/or structural analysis) in first grade (Vadasy & Sanders, 2011) and grades 2 through 3 (Vadasy et al., 2006B; 2007) generated average effect sizes at or above 0.33 SD. However, evaluations of a repeated reading intervention called *Quick Reads* showed smaller effect sizes of closer to 0.10 SD in second to third grade (Vadasy & Sanders 2008B; 2009) and 0.20 SD in fourth grade (Vadasy & Sanders, 2008C). Although relatively small-scale by impact evaluation standards, considering these high-quality studies together provides interesting insights into variation in effect sizes across grade levels with the principal investigator and many other treatment and study characteristics held constant. The pattern of declining returns for reading programs generally holds within this series of studies.

Lee et al. (2011) evaluate Experience Corps (EC), a nonprofessional tutoring program that uses "older adults" as reading tutors, with a sample of nearly 900 students across 23 schools in three cities. Run by the AARP foundation (https://www.aarp.org/experience-corps/), Experience Corps includes around 2,000 tutors and 20,000 students spread over 23 cities (Lee et al. 2011, p 98). Lee et al.'s (2011) evaluation includes more than 800 students in 23 schools with EC programs in Boston, New York City, and Port Arthur, Texas, evaluated over the 2006-2008 academic years. EC program affiliates employ paid staff members to recruit and train tutors, and to oversee program activities. Tutors receive between 15 and 32 hours of training. Different sites may choose their own locally-relevant curricula. The program runs for a full academic year, with two to four sessions per week; each session lasts about 30 to 40 minutes (p. 102). Outcomes included three standardized reading tests. Effect sizes were greater by 0.13-0.17 SD among students who received at least 35 sessions (p. 110). Results were robust across most subgroups, although effect sizes were higher for students in New York than the other cities, and non-IEP students benefited more than IEP students on one of the three standardized reading tests (p. 111).

Perhaps the most noteworthy math intervention for PreK-kindergarten is ROOTS, which uses paraprofessional "instructional assistants" hired and supervised by the school district as tutors. Effect sizes range from modest (at 0.10 SD) to substantial (0.57) SD but are consistently positive (Clarke et al. 2016A; 2017; Doabler et al. 2017) and particularly impressive, given the evident general difficulty in generating large effect sizes for early elementary math. Mattera et al. (2018) evaluate the High 5s program, a small-group kindergarten math program using tutors hired by a nearby teaching college. The effect of High 5s appears to have been more modest than ROOTS, remaining below 0.20 SD in our calculations. But it is worth noting that this program was layered atop a PreK math curriculum change and that High 5s at least generated a statistically significant

impact (relative both to students who received the curriculum change and pure control students) while the wider curriculum change did not generate any significant effects.

A first-grade small-group math tutoring intervention known as Number Rockets was evaluated at a small scale (Fuchs et al., 2005), at a larger scale (Fuchs et al., 2013, referred to here as Galaxy Math), and in a full-scale multistate impact evaluation in Gersten et al. (2015). The first two used graduate students as tutors, while the scaled-up evaluation used school-employed paraprofessionals. Gersten et al.'s (2015) study involves a sample of nearly 1,000 students in 76 schools across four urban school districts. They explicitly set out to bridge the gap between smaller-scale efficacy trials and larger-scale evaluations. Tutoring was delivered in small groups of two to three students per tutor. Scheduling was arranged so that students would not miss their normal math classes. Tutoring lasted for about 17 weeks, with at least three tutoring sessions held per week to meet a goal of 45 lessons in total, lasting about 40 minutes each.

Parker et al. (2019) evaluate a math intervention targeted towards Grades 4-8. The study included about 500 students in 13 schools across Minnesota (the program is available in more than 150 schools in Minnesota (p. 397). Tutors were "community members" who had made a year-long commitment to tutoring as part of the AmeriCorps program. Tutors received four days of training before the program, and two monthly two-hour follow-up sessions from doctorate-level practitioners, along with monthly meetings with a coach (p. 400). Tutoring was given for 90 minutes per week, divided into two or three sessions (p. 399). The intervention lasted for one semester, with outcomes measured in the winter. The authors report an effect size of 0.17 on the STAR math standardized assessment, with the effect size increasing to 0.24 SD under "optimal dosage conditions."

Cook et al. (2015) report on the only major high school tutoring intervention included in our meta-analysis in an evaluation with a sample of over 2,700 male students in grades 9-10 across 12 Chicago public high schools during the 2013-2014 school year. Students in the sample are overwhelmingly black or Hispanic (95%) and eligible for free- or reduced-price lunches (90%). Program administrators hire recent college graduates who are not certified teachers, but who commit to working as tutors for a year while receiving a small stipend. Tutoring occurs in 55-minute sessions daily that are organized into students' class schedules and extend for a full school year, with one tutor and two students in each session. The program generated effect sizes of 0.19 to 0.31 SD on standardized math test scores. Although these effect sizes are far from the largest found in our sample, they are exceptional relative to potential alternatives at the secondary level.

Program delivery

We turn next to program delivery characteristics. In aggregate, the pooled effect size for during-school tutoring programs (roughly one standard deviation) is nearly twice as large as that of after-school tutoring programs (roughly two-fifths of a standard deviation). During-school versus after-school variation occurs entirely within the subsample of paraprofessional and nonprofessional tutoring programs, since there were no after-school teacher tutoring programs and only one during-school parent tutoring program in our sample. The point estimate for the effect size of after-school nonprofessional tutoring programs is higher than that of during-school nonprofessional programs, but this finding should not be overstated given the relatively few after-school nonprofessional tutoring programs in our sample.

Table 3B indicates that pooled estimates for during-school interventions are higher than those for after-school interventions in all grade-level categories except for grades 6-11, for which

we do not have a large enough sample to interpret. Nearly all nonprofessional and parent tutoring programs had a one-to-one tutoring ratio. For teacher and paraprofessional tutoring, one-on-one tutoring showed the largest effect sizes, with ratios above one-to-one statistically similar. Grade-level dynamics in Table 3B appear in general to overshadow tutor-student ratio in moderating effect sizes, although interventions with ratios of three students per tutor and above seem to perform especially well in grades 2-5.

Finally, Table 3A indicates that effect sizes increase positively with the number of tutoring sessions per week. However, Table 3B shows that differences between 3 and 4-5 days per week are explained by preschool through grade 1 estimates, whereas grades 2-5 show higher effects for 3 days per week than for 4-5 days per week. There is little evidence of once-weekly tutoring sessions generating large effect sizes. In one noteworthy progression, Miller & Connolly (2013) find no significant effects from a weekly reading tutoring program for 8-9 year olds in Northern Ireland with nonprofessional tutors recruited through a business network. However, Miller et al. (2012) find significant (albeit modest) effects from the same program administered twice per week. Ritter's (2000; Ritter & Maynard 2007) lack of significant findings for a seemingly well-designed and implemented program may stem in part from the program's reliance on once-weekly tutoring sessions.

Counterintuitively, intervention duration of longer than 20 weeks show a pooled effect size that is slightly smaller than longer-term interventions. However, this relationship may be an artifact of the tendency for teacher tutoring programs to have relatively short durations, while nonprofessional tutoring programs tend to have longer durations. Moreover, treatment duration might be expected to shape longer-term outcomes more than shorter-term outcomes, and our sample allows us to analyze only the latter group in the meta-analyses.

Multivariate metaregressions

[Table 4 about here]

Table 4 shows results from regressions following the same model types as those in Tables 2, 3A, and 3B, except with covariates added in an attempt to begin quantitatively disentangling causal dynamics. The first row shows overall pooled effect sizes from all studies and estimates as a baseline, along with standard deviations calculated by averaging estimate standard deviations by treatment arm and then taking the mean of these standard deviations. Model 1 controls exclusively for study characteristics, Model 2 adds variables for sessions per week and intervention duration, and Model 3 adds dummies for math focus and after-school delivery. Finally, Model 4 adds tutor type controls and Model 5 adds dummy variables for a ratio of two or more students per tutor and for grade level (with first grade as the reference group).

Looking across models, it is immediately apparent that few covariates register as statistically significant according to conventional p-value thresholds, and differences tend to be moderate at best in terms of education significance. *We believe this trend is evidence of the robustness of tutoring programs across a wide array of contextual factors.* While the specifics of the estimates are substantially influenced by small study and estimate sample sizes in some sub-categories, changes are relatively minor across the specifications presented in Table 4.

The three study characteristic covariates are included in all models, but only the natural logarithm of sample size is statistically significant at all, and this variable is statistically significant in four of the five specifications, albeit marginally. As explained above, the negative association between sample size and impact magnitude is explained largely by nonprofessional tutoring programs, so this association does not call our broad meta-analytic findings into question. The

sample size coefficient loses its statistical significance and almost all of its magnitude once nonprofessional tutoring interventions are dropped. However, the magnitude and especially the consistency of the log sample size coefficient lend weight to the notion that the nonprofessional tutoring programs that have been randomly evaluated thus far on relatively large scales show relatively small impacts. This support strengthens our confidence in our finding that teacher and paraprofessional tutoring programs have shown more promise up to this point than have nonprofessional tutoring programs.

Models 4 and 5 include dummies for paraprofessional, nonprofessional, and parent tutoring, with teacher tutoring as the reference group. The negative signs on all three coefficients for both models show that teacher tutoring scores highest in terms of point estimates, even controlling for other characteristics. However, the significance of the coefficients is largely absorbed by the other covariates. The coefficient for having student-tutor ratios of two or more is negative but nonsignificant. Grade level dummies show that, even net of other covariates, grade level seems to vary inversely with impact size, notwithstanding an insufficient sample size for judgement above grade 5. The differential between grades 1 and 2-5 is negative and statistically significant, although of relatively small magnitude.

Discussion

This section locates the findings outlined above within their broader policy and research contexts. To start, we place our findings in perspective by considering the study's limitations. We then consider implications for scaling up tutoring interventions, focusing on 12 recent large-scale

impact evaluations. Finally, we suggest areas for future research and policy experimentation highlighted by our findings.

Limitations

This article faces some limitations that should be considered when interpreting the results. First and most importantly, this study faces a limitation faced by all meta-analyses dealing with diverse program designs and samples: our findings are dependent on insights regarding those programs that have been evaluated. Nonetheless, in comparison with other recent meta-analyses of education interventions, we have a large study sample size with consistent methodology (given our focus on RCT evidence) and a well-defined program model. To compare our study with two recent comparable others, Dietrichson et al. (2017)[7] included 101 studies, but the studies used a variety of different methodologies beyond RCTs and they included evaluations of a wide range of program models with bearing on low-SES students. Conn (2017) includes 56 articles that study education interventions in Sub-Saharan Africa, experimental and quasi-experimental, and the analysis includes 12 distinct intervention types. We draw inspiration from these authors' strategic use of the available data to draw the best inferences possible given existing research and contextual information.

Second, curriculum and other pedagogical characteristics of tutoring interventions remain mostly black-boxed in our review, except where compelling anecdotal evidence emerges. High-quality experimental studies abound in educational psychology and cognate disciplines on the specific pedagogical underpinning used by tutoring interventions. While eligible studies in this vein are included in our review, differences in curriculum were too subtle and multifaceted for us

[7] This study was used as a model of meta-analysis best practices in a recent *Review of Educational Research* methodology paper (Pigott & Polanin, 2019).

to code and quantitatively analyzed. Although we considered pedagogical characteristics wherever possible in the narrative review, we were not able to provide them with more than a cursory discussion. We felt this sacrifice was necessary given this project's central goals of developing a unified framework for analyzing tutoring interventions with the goal of building evidence for scaleup.

Third, the programs we tested typically focused on students who had fallen behind their respective grade levels and, in many cases, such students would be given alternative supplementary services—which often means some form of tutoring. These non-tutoring services vary across studies, and the extent to which they are measured and discussed within evaluations vary. While there is no complete solution to this problem, we paid careful attention to services given to control group students when they were reported, and we discuss such services in the text where relevant. Furthermore, to the extent that this issue does lead to bias, it would most likely lead to under- rather than over-reporting effect sizes.

Despite these shortcomings, the experimental literature on tutoring that has accumulated over the past three decades offers a treasure trove of data with a balance of consistency and diversity. We feel that PreK-12 tutoring interventions were long overdue for a comprehensive meta-analysis.

PreK-12 tutoring programs at scale

In this section, we focus on experimental evidence on tutoring programs implemented and studied at relatively large scales. A central contribution of this review is to integrate insights from the handful of recent large-scale impact evaluations with those arising from the dozens of smaller-scale efficacy trials that have been carried out over the past four decades. Smaller-scale programs

may circumvent some of the program management and oversight challenges that afflict larger-scale programs, although, within some ranges, economies of scale may emerge. Evaluating tutoring programs within smaller or more homogeneous samples generally allows for stronger control over program implementation and thus more precise estimates of how the program performs when it operates as planned. Efficacy trials may be especially well-suited to test alternative curricula and pedagogical practices, since researchers are likely to have more control over program implementation.

However, administrative complexities and other difficulties associated with maintaining program fidelity are endemic to the education policy environment and must be accounted for. As Gersten et al. (2015) point out, there is "a loose coupling between often precise theoretical underpinnings of the best efficacy trials and the broad, often eclectic theoretical underpinnings of large-scale federal, state, or local initiatives…implementation is often carefully monitored in the controlled efficacy trials but allowed to vary widely in most of the large-scale evaluation studies" (p. 517). "Contemporary thinking about large-scale evaluations argues for a combination of efficacy trials (conducted in controlled settings) to test whether an intervention can produce a significant impact on important outcomes, followed by a series of much larger, less tightly controlled scale-up studies to test whether an intervention can work in the real world of typical school settings" (p. 518). The present section focuses on recent studies with relatively large sample sizes—in particular, the 12 studies included in our systematic review that use samples of ~400 or more students and that have been carried out since 2010. Table 5 lists key study and program characteristics associated with these evaluations.

As Table 5 demonstrates, large-scale impact evaluations show substantial effects of generally comparable magnitude to efficacy trials and other smaller-scale studies. Taken

collectively, the studies outlined in Table 5 indicate that tutoring programs can exert strong impacts for a wide range of samples and over diverse intervention characteristics. Math interventions are better represented within the sub-sample of large-scale impact evaluations than within our study sample as a whole. Impacts appear relatively even across subjects and grades, notwithstanding the small number of evaluations in the post-elementary grades.

The type of tutor employed may become especially important when scaling up activities. While plausible interaction effects between intervention scale on one hand and subject area or grade level on the other do not readily come to mind, the supply of tutors who meet particular qualifications may constrain scaleup. If implementers must tap tutor pools other than those considered optimal for the study, effect sizes may fall. Whether this fall occurs for a particular program depends on the nature of the tutor supply and on the robustness of the tutoring program content to tutors with different characteristics. Davis et al. (2017) formalize a generalization of this intuition, arguing that implementers rank the quality of available inputs and select inputs of the highest quality. In tutoring programs, as in many other social interventions, human skills and agency constitute key inputs. Within this framework, the most efficiently scalable programs will be those that are most robust to differing tutor characteristics at the margins.

Not enough recent large-scale impact evaluations have been carried out to support strong assertions about how the four tutor type categories' impact varies at larger scales. However, Table 5 does show that teacher tutoring, paraprofessional tutoring, and nonprofessional tutoring programs can all generate large effect sizes, even when scaled up. The majority of studies in our impact evaluation subsample are concentrated within the paraprofessional (six studies) and nonprofessional (four studies) categories. No parent tutoring studies fit the criteria.

Only two recent large-scale randomized studies have evaluated teacher tutoring interventions (Sirinides et al. 2018; Smith et al., 2013). The relative scarcity of these evaluations may arise because "tutoring by regular teachers is widely viewed as too costly to undertake on a large scale" (Ander et al., 2016, p. 4). In practice, Reading Recovery has been one of the sole teacher tutoring programs that has been widely implemented and widely evaluated. It is no surprise then that the two program models recently evaluated on a large scale are Reading Recovery (Sirinides et al., 2018) and its counterpart, Mathematics Recovery (Smith et al., 2013). Both of these interventions showed substantial effects—especially Reading Recovery—but studies with other tutor types suggest that it may be possible to achieve comparable effects with lower tutor costs. We next compare programs designed for different tutor types for reading, before moving on to math.

Sirinides et al.'s (2018) large-scale evaluation of Reading Recovery shows the strongest average impact of a reading intervention in the table, with an effect size of nearly 0.5 SD. The magnitude of this effect size falls centrally in the range of those found by the smaller-scale studies on Reading Recovery discussed above, lending weight to the robustness of Reading Recovery's effects to a wide range of study samples and implementation contexts. To the extent that an intervention model can consistently prove its ability to exert strong effects on first-grade reading in many places through research studies, Reading Recovery has done so. However, it is less clear that even Reading Recovery can substantially outperform well-designed paraprofessional reading programs, particularly given the cost reductions associated with paraprofessional relative to teacher tutoring programs, and the fact that first-grade reading programs on average tend to see relatively large effect sizes compared to other subject-grade combinations.

The Minnesota Reading Corps (MRC) program represents a case in point, although it is worth noting that comparability with Reading Recovery for present purposes is limited by the fact that MRC operates in only a single state, in contrast to the Sirinides et al. (2018) study which evaluated Reading Recovery across the United States. In a recent NORC evaluation, Markovitz et al. (2014) found that MRC increased first-graders' reading scores by 0.37 SD, falling only slightly short of Reading Recovery's impact of nearly 0.5 SD, presumably at substantially lower costs. At 1.06 SD, MRC's impact on kindergarten reading was among the largest of any estimates in our study, while effects were small in Grade 3 and non-significant in Grade 2. These patterns may indicate that grade level outweighs subtleties in tutoring pedagogy in accounting for effect sizes.

Among nonprofessional literacy tutoring programs, impacts found by Jacob et al. (2016), Lee et al. (2011) and Miller et al. (2013) are relatively modest in the range of 0.10 – 0.20 SD, but impressive nonetheless, considering the tutors are unpaid and receive minimal training and supervision. The sole study not to find a significant impact was Miller & Connolly's (2013) evaluation of Time to Read, and the version evaluated in this study consisted of only 30 minutes of tutoring once per week. Miller et al. (2012) find significant (albeit modest) effects for an adjusted version of the program that included additional training and held tutoring twice per week.

Turning to math interventions, Smith et al. (2013) find that Mathematics Recovery exerted a meaningful impact on learning outcomes, although with smaller effect size magnitudes than Reading Recovery. Standard deviations may be difficult to compare across subject areas, but Mathematics Recovery's effect sizes are also lower relative to paraprofessional math alternatives. In particular, Number Rockets shows strong promise. In the largest-scale evaluation to date, Gersten et al. (2015)'s estimated effect size of 0.34 SD on a standardized math test score matches findings from Fuchs et al.'s (2005) smaller-scale trial, and falls toward the center of the range of

findings reported in the program variations evaluated in Fuchs et al. (2013). This finding is especially noteworthy, given that the version of the program implemented in Gersten et al. (2015) drew on a much larger and more diverse sample across four states, and employed community members, rather than students, as tutors (p. 523). The two other paraprofessional pre-secondary math studies found effect sizes between 0.10 and 0.20 SD for a kindergarten program with school employees as tutors (Mattera et al. 2018), and Grades 4-8 with AmeriCorps service fellows as tutors (Parker et al. 2019).

The findings from Cook et al.'s (2015) math tutoring program are among the most noteworthy in the study. The program model they evaluate was originally developed for use at MATCH Charter Public High School in Boston and is now administered in multiple states by the nonprofit organization Saga Education. The model rests on five main characteristics: daily tutoring sessions; in-school delivery; personalized instruction; supportive relationships with near-peer tutors; and an evidence-based curriculum.[8] The authors find effect sizes of roughly 0.20-0.30 SD for standardized math test scores and 0.50 SD for grades. Although the impact on standardized test scores are not among the largest in effect size magnitude, this may in part arise from greater variation in high school, leading to large standard deviation. Perhaps more importantly, these effect sizes are impressive, given the long-noted greater difficulty in generating large effect sizes for older vs. younger children (Carneiro & Heckman, 2003). Most importantly, effects appear large enough to be potentially transformative at low costs:

> "Match/SAGA tutorials helped students learn between one and two extra years of math above what the typical American high school student learns in one year…tutorials moved students on average from about the 34th percentile to about the 42nd percentile in the national distribution…[closing] about half the gap between participants' math scores prior to the tutorials and the national average" (Ander et al. 2016).

[8] https://www.sagaeducation.org/our-approach

Cook et al.'s (2015) paper presents the only publicly available results with Saga tutoring tested on its own. However, Cook et al. (2014) also find large effects of the program in a pilot study, in which some treatment students also received a group-based social cognitive program for 9^{th}-10^{th} graders in Chicago. Fryer et al. (2014) find large impacts from a multi-component program implemented in Houston Public Schools (grades 3-9) that includes math tutoring following the Saga model. Nonexperimental comparisons in this study between treatment schools that received tutoring and those that did not indicate that the tutoring likely exerted a substantial independent effect above and beyond the other treatment components (p. 1389).

The Saga tutoring model was developed for mathematics, but some experimentation has begun with adapting the model for literacy tutoring. While no experimental results have yet been released that isolate the effects of reading tutoring following the Saga model, two recent studies rigorously evaluate these reading programs and find some evidence of potential. A small-group reading tutoring program following a similar model in New York City public middle schools did not show evidence of strong effects (Fryer & Howard-Noveck, 2020), in line with previous evidence that math programs tend to generate larger effects than reading programs at the middle school level (p. 422). However, the study did find a meaningful effect of just under 0.10 SD for black students. This study was not included in our meta-analysis since treatment students received an after-school program of which tutoring was only one component, so tutoring effects cannot be separated out using their experimental estimates. Additionally, Kraft (2015) finds effect sizes of 0.15-0.25 SD for literacy outcomes and no significant results for math among tenth-graders in Boston using a quasi-experimental analysis.

Contextualizing PreK-12 tutoring programs

How do the strategies and impacts discussed above compare to alternative PreK-12 interventions? In this section, we consider the findings discussed above within the context of broader education policy debates on tutoring and on comparable interventions that may potentially complement or substitute for tutoring programs. We begin by noting the findings of recent meta-analyses that have analyzed the effects of tutoring alongside other learning interventions; we then summarize recent findings on peer and cross-age tutoring, mentorship, and computer-assisted learning, and consider our findings in light of this wider field.

As discussed above, Dietrichson et al.'s (2017) meta-analysis of treatment-control group design studies yielded a pooled effect size for tutoring programs of 0.36 SD, nearly identical to our estimate. To place this effect size in perspective, their analysis estimated effect sizes of 0.32 SD, 0.24 SD, and 0.22 SD respectively for "feedback and progress monitoring," "small-group instruction" (which differs from tutoring in that it involves groups of six or more students) and "cooperative learning." While 36 studies are included in the tutoring estimate, the other three estimates are based on findings from only five, four, and ten studies respectively. The remaining ten intervention components that their meta-analysis includes show pooled effect sizes below 0.20 SD (Dietrichson et al, 2017, p. 268).

A series of "best-evidence synthesis" (Slavin, 1986) reviews that examine the effectiveness of a variety of educational programs across the literacy versus math divide adds further perspective. Pellegrini et al. (2018) review eight categories of elementary math interventions evaluated across 78 studies, and find that tutoring programs have by far the highest average effect sizes. These effect sizes are also similar to our pooled findings, with 0.26 SD for one-on-one tutoring and 0.32 SD for small-group tutoring programs. Of the other categories, only

"instructional process programs" came close at 0.25 SD, with the rest of the categories showing average effect sizes of under 0.10 SD.

Inns et al. (2019) find average effect sizes of 0.31 SD for one-on-one tutoring and 0.20 SD for small group tutoring interventions in primary school. These scores are also high relative to the other categories studied, but non-technology school- and classroom-level also show substantial positive influence, leading the authors to recommend including tutoring along with, for instance, curricular developments and collaborative learning programs in educational reforms. Baye et al. (2018) come to similar conclusions by reviewing the literature for reading interventions at the secondary school level.

It is thus clear that tutoring programs tend to perform well when evaluated alongside comparable intervention. But what have researchers found thus far about the effects of programs that share key characteristics in common with tutoring? We begin with *peer* and *cross-age* tutoring which, as explained above, are often referred to as tutoring but which are excluded from our meta-analysis for pragmatic, analytical reasons. Peer tutoring generally refers to programs in which children of the same age or in the same grade tutor one another. Similar to the present review, Dietrichson et al. (2017) consider peer tutoring to be a form of "cooperative learning" rather than tutoring as usually understood, and find an average of 0.22 SD from evaluations of interventions falling into this category. Cross-age tutoring refers to situations in which older students tutor younger students, but there has been less rigorous research conducted on cross-age than peer tutoring. The research that has been conducted suggest that these have similar effects as peer tutoring, with one recent meta-analysis pooling cross-age and peer tutoring separately and finding effect sizes of 0.26 SD for both (Slavin et al., 2009, p. 1442).

Peer and cross-age tutoring may reduce costs relative to programs that require paid tutors, and could generate positive spillover effects to the extent that tutoring benefits the tutor as well as the tutee. However, on the negative side, it may be difficult to ensure consistently high-quality tutoring from children, and the ethical necessity of ensuring benefit to the tutor as well as the tutee may present logistical difficulty. While a great deal remains to be learned, existing research suggests that peer and cross-age learning programs may be highly effective, but are unlikely to substitute to any substantial extent for programs that employ adults as tutors.

Peer-Assisted Learning Systems (PALS),[9] the most prominent peer tutoring program model in the US, was designed by Lynn and Douglas Fuchs and collaborators at Vanderbilt University (in conjunction with local school teachers) and has been the subject of numerous studies beginning in the 1990s (Fuchs et al. 1995; 1997). The PALS model builds on Class-Wide Peer Tutoring, a model first used during the 1980s in which entire classrooms are broken up into pairs or small groups of students and students tutor one another (Delquadri et al., 1986). Since its launch, PALS has been adapted for use across all grade levels from Pre-K through secondary school, and for both reading and math. PALS sessions occur during school, typically during a time devoted to the relevant subject (i.e., math PALS during math time and reading PALS during reading time). Sessions usually occur for 30-45 minutes, around three times per week. The teacher ranks students on strength with regard to the relevant skills, divides the distribution in two, and then pairs the top student in the top half with the top student on the bottom half and so on, in order to ensure that there is one student who is relatively strong in the subject matter is placed within each pair. The teachers move from pair to pair, observing and providing feedback. Students typically switch at some point, making the tutoring "reciprocal" (McMaster & Fuchs, 2016).

[9] https://frg.vkcsites.org/what-is-pals/

For the most part, PALS efficacy trials have tended to show consistently strong effects, some even exceeding a full standard deviation for certain subgroups (Fuchs and Fuchs, 2005). There have been fewer large-scale evaluations or studies that extend outside of the Nashville area where PALS originated. One recent large-scale impact evaluation conducted in Minnesota, Tennessee, and Texas found significant overall impacts ranging from 0.29 to 0.42 SD (Stein et al., 2008). However, effects were substantially stronger in Tennessee, where the program was developed and has the strongest supportive infrastructure.

Overall, it is clear that PALS is highly effective at improving skills across multiple grades and subjects within its core geographic area, and could likely expand these further if adaptation to new contexts continues. However, we see the PALS model and other peer tutoring initiatives as an important step toward the customization of classroom learning, rather than a supplementary intervention as in the tutoring programs included within our meta-analysis. Even in contexts where PALS is used within classrooms, some students can likely benefit from additional adult tutoring.

Mentoring, like peer tutoring, contains substantial areas of overlap of tutoring as understood in the present article. Like tutoring, mentoring involves the cultivation of a relationship between a student and a model figure but, unlike tutoring, is not primarily based around educating the student on specific academic tasks. Dietrichson et al. (2017) find a non-significant pooled impact of 0.04 SD on student mentoring programs. A recent evaluation of a high school mentoring program in France with a sample of roughly 500 students finds that, although mentored students reported high satisfaction with the intervention, small negative effects on learning outcomes arose. The authors attribute this counterintuitive result to a time tradeoff, in which the mentoring activities are less productive than homework time. This finding may also strike a cautionary note for after-school tutoring programs, particularly at higher grade levels where homework figures

more heavily into the students' education (Ly et al. 2020).[10] However, given that lack of specific academic focus distinguishes mentoring from tutoring as the terms are commonly used, this lack of evidence for impact on learning outcomes is not surprising. Many mentoring programs have been found to reduce delinquency and related behaviors thought to negatively impact learning (Tolan et al., 2013). These benefits could in turn translate into positive effects on longer-run learning trajectories.

Finally, many computer-assisted learning programs (CAL) are thought to emulate elements of tutoring programs, so much so that responsive learning programs have come to be known as "intelligent tutoring systems" (Ma et al., 2014). Like tutoring programs, CAL programs can be used as a substitute for classroom time or other activities during the school day, or can be completed after school, replacing time that would otherwise be spent on homework or extracurricular activities. To the extent that computer programming methods are able to approximate the instructional feedback that human tutors would otherwise provide, CAL programs could represent a less costly alternative to human tutoring. However, the lack of human engagement may remove some of the potential benefits of tutoring, including associating positive human interaction with the educational content. These dynamics may vary across age and developmental stage. In a recent meta-analysis of intelligent tutoring systems, Fletcher & Kulik (2017) find an impressive pooled effect size of 0.66 SD. However, this analysis includes non-experimental as well as experimental studies, and also lab studies which may be more likely to show large effect sizes than field-based studies. In a systematic review of experimental education technology research, Escueta et al. (2017) show that the majority of randomized evaluations of

[10] The authors refer to the intervention evaluated as "tutoring," but it does not meet our definition of tutoring since tutors are generally asked to help the students prepare for postsecondary education, rather than assigned to teach specific content.

CAL programs show positive effects and that math programs tend to show particularly large effect sizes.

Implications for theory, policy, and future research

While a great deal remains to be learned with regard to the optimization of tutoring programs, the present article presents results that robustly support several theoretical and practical propositions. In this section, we consider these propositios in the context of today's education policy environment, and consider how the insights we synthesized from the experimental literature can best be leveraged within the contemporary policy environment.

First, we feel that there is immense scope for exploration, development, and implementation of paraprofessional tutoring programs in particular. As pointed out in Cook et al.'s (2015) explication of the Saga Education math tutoring model, the skills required for effective tutoring are distinct from the skills required for effective classroom teaching. And while teachers may often make effective tutors, it is far from clear that effectiveness differentials between trained teachers and paraprofessionals outweigh the cost differentials. We certainly do not advocate retrenchment of Reading Recovery—that particular program has honed its approach over decades, has built up administrative infrastructure that allows for economies of scale, and has demonstrated cost-effective results. However, we feel that paraprofessional tutoring represents a much more expansive area for growth in tutoring programs given the clear potential for transformative effects at relatively low costs, even for high-dosage and/or one-on-one tutoring.

Several nonprofessional tutoring programs have shown promising results. As discussed above, this category consists primarily of unpaid volunteers. Many of the program models that have been evaluated arose from a push that began during the 1990s for more volunteering

surrounding education. Where suitable volunteer pools are available, programs utilizing them as tutors are likely to prove productive. But it is less clear that volunteers represent a suitable pool of tutors for scaled-up programs and within broader policy planning, since these programs typically allow less scope for training and dedicated commitment. The experimental parent tutoring research is still too thin and fragmented for consistent lessons. As is the case for nonprofessional tutoring, our review highlights several programs that yielded large effects given ostensibly low costs. However, program designers typically have much less control over parent tutoring implementation than over the other types. And although we believe that training and deploying parents to tutor children is likely to be productive in many cases, these programs might be best integrated with broader family support services. Paraprofessional tutoring thus seems to us to be the priority area for future tutoring planning.

In particular, paraprofessional school staff members and recent graduates in professional fellowship programs represent promising bodies of potential tutors. As economies adjust to automation and undergo other structural changes that reduce the prominence of manufacturing, scholars and policy analysts have begun to plan for shifts toward the expansion of job markets in human services. Education represents a case in point. Tutoring and other in-school intervention activities may represent a viable and fulfilling career path for many individuals who might not otherwise enter the education sector. Programs that employ paraprofessional school staff members as tutors may save on administrative costs given their integration into the school, and may allow for stability as the programs develop. Relatedly, education-oriented civic programs are becoming increasingly common within the career trajectory of recent college graduates. AmeriCorps and other civic fellowship programs are likely to continue to create large pools of potential tutors for

many years to come. Future research should explore whether, and under which conditions, school staff members generate higher outcomes than post-graduate fellows, or vice versa.

Relatedly, we feel that there is a large scope for expansion of tutoring at the secondary level. While effect sizes may tend to be higher at the early elementary level than for higher grade levels, the most relevant point of comparison for policymakers with regard to a tutoring program for a particular grade level is typically the opportunity cost of that program relative to other intervention opportunities for that grade level (rather than the opportunity cost of investing in the grade level in question versus another one). Even if fewer tutoring evaluations show high-magnitude effects at the secondary level, they may be potentially transformative if other secondary intervention options give rise to even smaller effect sizes. The Saga Education model represents an especially promising administrative-pedagogical model for expansion. Similarly, experimental evaluations of tutoring programs in subject areas other than reading and math, for instance science or social studies, could open a new area for tutoring policy research. While it may make sense to focus on reading and math at earlier grade levels, as tutoring research expands at the secondary level it will become increasingly important to understand how effect sizes vary across a broader range of subjects.

In terms of program delivery, the panoramic view provided by our meta-analysis suggests that, at least when implemented on larger scales, the extent to which program implementers are able to ensure that tutoring actually occurs at sufficiently high doses may outweigh subtleties in the content being taught. While our data do not allow us to address this topic statistically, our review suggests that the relatively lower effects found within after-school and parent tutoring may arise largely from difficulties in ensuring that tutoring actually occurs as planned within these

contexts. Treatment on the treated estimates, if reported, would likely be much higher than intent-to-treat estimates, as is the case for Cook et al.'s (2015) findings.

As for any policy arena, researchers and practitioners must pay close attention to impacts on equity as well as on overall effect sizes. For the most part, it seems likely that tutoring programs following the models evaluated in this review would be on net equity-increasing. The vast majority of programs examined in the tutoring academic and policy literature are implicitly or explicitly conceptualized as remedial. Since lower academic scores tend to correlate with lower socioeconomic status, ethnic or racial marginalization, and other layers of structural disadvantage, remedial programs should lessen the education gap. However, some types of tutoring may be more equity-inducing than others, depending on their effectiveness for particular groups of students. Future studies should pay close attention, where possible, to whether and how effect sizes differ for different social groups. While the specific mechanisms explaining such divergences may not be obvious *a priori*, they may still exert powerful effects. For instance, despite Fryer & Howard-Noveck's (2020) findings of a modest effect size for middle school reading tutoring, effects for Black students were substantially larger, yielding one of the study's most noteworthy findings.

At a broader level, equity considerations may underscore the importance of free tutoring programs more broadly. Parents of privileged students, like education researchers and practitioners, have noticed the potential effectiveness of tutoring and have attempted to leverage it. Over the past few decades, private tutoring that households must pay for has grown increasingly popular. Its influence within education systems more broadly has grown to the point that private supplementary tutoring is sometimes referred to as a "shadow education system" (Bray 2013). Private tutoring may thus constitute a potential mechanism for "dream hoarding," the process by which upper- and even middle-class parents reproduce inequities by monopolizing human

development opportunities (Reeves, 2017). Increasing the presence and effectiveness of public tutoring systems may thus be important for less-advantaged students to keep pace in this environment.

Both equity and efficiency considerations further point toward the importance of identifying the populations of students who could most benefit from tutoring. In particular, students who have fallen behind as a result of structural contingencies rather than specific learning disorders may especially benefit from tutoring programs that can set them on self-sustaining pathways toward rapid learning. A relatively distinct literature has already emerged on tutoring for foster children. Future studies could also test interventions for other marginalized populations of children whose circumstances may have precluded sufficient preparation for regular school, including incarcerated adolescents (Wexler et al. 2014) and refugees (Naidoo 2008; 2009). And although a growing number of studies investigates the impact of tutoring on learning outcomes for ELL students, this area of research has enormous room for growth as well.

Conclusion

Tutoring programs rank among the most flexible and potentially transformative learning program types available at the PreK-12 levels. While this proposition has been clear for some time, the present review has, for the first time, synthesized and quantitatively analyzed experimental evidence on all programs for which such evidence is available that would be widely identified as tutoring. With effect sizes averaging at over a third of a standard deviation and impacts consistently significant across a wide range of program and study characteristics, our review's meta-analytic findings demonstrate not only the power of tutoring, but its versatility. As customized learning

grows in prominence across today's educational systems, there is little doubt that tutoring programs will constitute a key workhorse policy model.

Bibliography

Note: Bold citations refer to studies included in the meta-analysis.

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

Alegre, F., Moliner, L., Maroto, A., & Lorenzo-Valentin, G. (2019). Peer tutoring in mathematics in primary education: a systematic review. *Educational Review*, *71*(6), 767-791.

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.

Anania, J. (1981). *The effects of quality of instruction on the cognitive and affective learning…* Dissertation

Anania, J. (1983). The Influence of Instructional Conditions on Student Learning and Achievement. *Evaluation in Education: An International Review Series*, *7*(1), 3-76.

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.

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

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., & Walton, M. (2015). Teaching at the right level: Evidence from randomized evaluations in India. *NBER Working Paper*, *22746*.

Benner, G. J. (2004). An investigation of the effects of an intensive early literacy support program on the phonological processing skills of kindergarten children at-risk of emotional and behavioral disorders.

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.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, *13*(6), 4-16.

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* (No. 2019: 7). Working Paper.

Borenstein, M. (2009). Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis*, *2*, 221-235.

Borman, G. D., Borman, T. H., Park, S. J., & Houghton, S. (2019). A Multisite Randomized Controlled Trial of the Effectiveness of Descubriendo la Lectura. *American Educational Research Journal*, 0002831219890612.

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.

Burke, A. J. (1983). *Students' potential for learning contrasted under tutorial and group approaches to instruction* (Doctoral dissertation, University of Chicago).

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.

Center, Y., Wheldall, K., Freeman, L., Outhred, L., & McNaught, M. (1995). An evaluation of reading recovery. *Reading research quarterly*, 240-263.

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.

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).

Conn, K. M. (2017). Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. *Review of Educational Research*, *87*(5), 863-898.

Cook, J. A. (2002). Every moment counts: Pairing struggling young readers with minimally trained tutors.

Cook, P. J., Dodge, K., Farkas, G., Fryer Jr, R. G., Guryan, J., Ludwig, J., ... & Steinberg, L. (2014). *The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: results from a randomized experiment in Chicago* (No. w19862). National Bureau of Economic Research.

Cook, P. J., Dodge, K., Farkas, G., Fryer, R. G., Guryan, J., Ludwig, J., ... & Steinberg, L. (2015). Not too late: Improving academic outcomes for disadvantaged youth. *Institute for Policy Research Northwestern University Working Paper WP-15-01.*

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Davis, J. M., Guryan, J., Hallberg, K., & Ludwig, J. (2018). Scale-Up Experiments.

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.

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.

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.

Erion, R. J. (1994). Parent tutoring, reading instruction and curricular assessment.

Fives, A., Kearns, N., Devaney, C., Canavan, J., Russell, D., Lyons, R., ... & 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.

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.

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.

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.

Fuchs, L. S., Seethaler, P. M., Powell, S. R., Fuchs, D., Hamlett, C. L., & Fletcher, J. M. (2008A). Effects of preventative tutoring on the mathematical problem solving of third-grade students with math and reading difficulties. *Exceptional children*, *74*(2), 155-173.

Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., & Schatschneider, C. (2008B). 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.

Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., ... & 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.

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.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., ... & Bryant, J. D. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology*, *105*(1), 58.

Fuchs, D., Kearns, D. M., Fuchs, L. S., Elleman, A. M., Gilbert, J. K., Patton, S., ... & 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.

Fryer Jr, R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *The Quarterly Journal of Economics*, *129*(3), 1355-1407.

Fryer Jr, 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.

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.

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.

Goudey, J. (2009). A parent involvement intervention with elementary school students: The effectiveness of parent tutoring on reading achievement.

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.

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

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.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., . . ., Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, 343.

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 Encyclopedia, Center for Research and Reform in Education*.

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.

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.

Jung, P. G. (2015). Effects of data-based instruction for students with intensive early writing needs: A randomized control trial.

Lachney, R. P. (2002). Adult-mediated reading instruction for third through fifth grade children with reading difficulties.

Lam, S. F., Chow-Yeung, K., Wong, B. P., Lau, K. K., & Tse, S. I. (2013). Involving parents in paired reading with preschoolers: Results from a randomized controlled trial. *Contemporary Educational Psychology*, *38*(2), 126-135.

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.

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.

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.

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.

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.

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

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*.

Marquis, R. (2013). *The Gender Effects of a Foster Parent-Delivered Tutoring Program on Foster Children's Academic Skills and Mental Health: A Randomized Field Trial*. University of Ottawa (Canada).

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.

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. *New York: MDRC, March*.

Mayfield, L. G. (2000). The effects of structured one-on-one tutoring in sight word recognition of first-grade students at-risk for reading failure.

Mears, P. R. (2007). The Effects of the Fast Start Program on the Reading Achievement of Emergent and Beginning Readers: A Replication and Extension.

Mehran, M., & White, K. R. (1988). Parent tutoring as a supplement to compensatory education for first-grade children. *Remedial and Special Education*, *9*(3), 35-41.

Miller, B. V., & Kratochwill, T. R. (1996). An evaluation of the paired reading program using competency-based training. *School Psychology International*, *17*(3), 269-291.

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.

Mooney, P. J. (2004). 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.

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.

Naidoo, L. (2009). Developing social inclusion through after-school homework tutoring: a study of African refugee students in Greater Western Sydney. *British journal of sociology of education*, *30*(3), 261-273.

Nielson, B. B. (1992). Effects of parent and volunteer tutoring on reading achievement of third grade at-risk students.

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.

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.

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.

Pellegrini, M., Lake, C., Inns, A., & Slavin, R. E. (2018, October). Effective programs in elementary mathematics: A best-evidence synthesis. *Best Evidence Encyclopedia*.

Pigott, T. D., & Polanin, J. R. (2019). Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, 0034654319877153.

Pinnell, G. S., DeFord, D. E., & Lyons, C. A. (1988). *Reading Recovery: Early intervention for at-risk first graders*. Educational Research Service.

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*, 9-39.

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.

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.

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.

Powell-Smith, K. A., Stoner, G., Shinn, M. R., & Good III, R. H. (2000). Parent tutoring in reading using literature and curriculum materials: Impact on student reading achievement. *School Psychology Review*, *29*(1), 5-27.

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. *Literacy Research and Instruction*, *43*(4), 21-40.

Rasinski, T., & Stevenson, B. (2005). The effects of fast start reading: a fluency-basedhome involvement reading program, on the reading achievement of beginning readers. *Reading Psychology*, *26*(2), 109-125.

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

Rimm-Kaufman, S. E., Kagan, J., & Byers, H. (1998). The effectiveness of adult volunteer tutoring on reading among "at risk" first grade children. *Literacy Research and Instruction*, *38*(2), 143-152.

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*.

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research*, *79*(1), 3-38.

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.

Shenderovich, Y., Thurston, A., & Miller, S. (2016). Cross-age tutoring in kindergarten and elementary school settings: A systematic review and meta-analysis. *International Journal of Educational Research*, *76*, 190-210.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology*, *70*, 747-770.

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.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational researcher*, *15*(9), 5-11.

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.

Stein, M. L., et al. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis*, 30(4), 368-388.

Saenz, L., Yen, L., Fuchs, L. S., & Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis*, 30, 368–388.

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.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research synthesis methods*, *5*(1), 13-30.

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.

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.

Vadasy, P. F. & Sanders, E. A. (2008A). 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.

Vadasy, P. F. & Sanders, E. A. (2008B). Repeated reading intervention: Outcomes and interactions with readers' skills and classroom instruction. *Journal of Educational Psychology*, *100*(2), 272.

Vadasy, P. F. & Sanders, E. A. (2008C). Benefits of repeated reading intervention for low-achieving fourth-and fifth-grade students. *Remedial and Special Education*, *29*(4), 235-249.

Vadasy, P. F. & Sanders, E. A. (2009). Supplemental fluency intervention and determinants of reading outcomes. *Scientific Studies of Reading*, *13*(5), 383-425.

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*, 786.

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.

Vadasy, P. F., Jenkins, J. R., Antil, L. R., & Wayne, S. K. (1997A). Community-based early reading intervention for at-risk first graders. *Learning Disabilities Research & Practice*.

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.

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.

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.

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.

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.

Vaughn, Sharon, et al. "Effectiveness of an English intervention for first-grade English language learners at risk for reading problems." *The Elementary School Journal* 107.2 (2006): 153-180.

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.

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.

Wanzek, J., & Roberts, G. (2012). Reading interventions with varying instructional emphases for fourth graders with reading difficulties. *Learning Disability Quarterly*, *35*(2), 90-101.

Wolff, U. (2011). Effects of a randomised reading intervention study: An application of structural equation modelling. *Dyslexia*, *17*(4), 295-311.

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*. Rutgers The State University of New Jersey-New Brunswick.

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.

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.

Table 1
Number of Studies Used in Meta-Analysis and Proportion of Total Studies Used Arranged by Type of Study and Type of Treatment

Type of Study	Type of Treatment						
	All	Literacy	Math	Teacher tutoring	Paraprof tutoring	Nonprof tutoring	Parent tutoring
All	96 [1.00]	74 [0.77]	27 [0.28]	18 [0.19]	46 [0.48]	23 [0.24]	11 [0.11]
Subject							
Literacy	..	..	..	16 [0.17]	26 [0.27]	23 [0.24]	11 [0.11]
Math	..	..	..	3 [0.03]	21 [0.22]	2 [0.02]	1 [0.01]
During vs. after school?							
During	79 [0.82]	58 [0.60]	23 [0.24]	18 [0.19]	44 [0.46]	18 [0.19]	1 [0.01]
After	18 [0.19]	17 [0.18]	4 [0.04]	0	2 [0.02]	6 [0.06]	10 [0.10]
Grade/Level							
Presch-Kind	18 [0.19]	12 [0.13]	6 [0.06]	1 [0.01]	10 [0.10]	5 [0.05]	2 [0.02]
Grade 1	46 [0.48]	40 [0.42]	9 [0.09]	10 [0.10]	20 [0.21]	13 [0.14]	3 [0.03]
Grades 2-5	50 [0.52]	41 [0.43]	14 [0.15]	9 [0.09]	20 [0.21]	16 [0.17]	7 [0.07]
Grades 6-11	7 [0.07]	5 [0.05]	5 [0.05]	0	3 [0.03]	3 [0.03]	1 [0.01]
Tutor to student ratio							
1 to 1	67 [0.70]	58 [0.60]	13 [0.14]	13 [0.14]	22 [0.23]	22 [0.23]	11 [0.11]
1 to 2	14 [0.15]	8 [0.08]	6 [0.06]	2 [0.02]	12 [0.13]	1 [0.01]	0
1 to 3 or more	24 [0.25]	14 [0.15]	11 [0.11]	4 [0.04]	19 [0.20]	1 [0.01]	0
Special populations							
English lang learners	10 [0.10]	7 [0.07]	3 [0.03]	3 [0.03]	6 [0.06]	1 [0.01]	0
Foster students	3 [0.03]	3 [0.03]	3 [0.03]	0	1 [0.01]	1 [0.01]	1 [0.01]
Risk of bias							
Low	84 [0.88]	62 [0.65]	27 [0.28]	18 [0.19]	45 [0.47]	17 [0.18]	5 [0.05]
High	12 [0.13]	12 [0.13]	0	0	1 [0.01]	6 [0.06]	6 [0.06]

(continued)

Table 1 (continued)
Number of Studies Used in Meta-Analysis and Proportion of Total Studies Used
Arranged by Type of Study and Type of Treatment

Type of Study	Type of Treatment						
	All	Literacy	Math	Teacher tutoring	Paraprof tutoring	Nonprof tutoring	Parent tutoring
Student sample size							
50 or fewer	22 [0.23]	20 [0.21]	3 [0.03]	3 [0.03]	6 [0.06]	9 [0.09]	5 [0.05]
50 to 100	24 [0.25]	22 [0.23]	5 [0.05]	5 [0.05]	9 [0.09]	7 [0.07]	3 [0.03]
100 to 200	23 [0.24]	17 [0.18]	6 [0.06]	5 [0.05]	16 [0.17]	0	2 [0.02]
200 to 400	12 [0.13]	8 [0.08]	5 [0.05]	3 [0.03]	7 [0.07]	2 [0.02]	1 [0.01]
400 and up	15 [0.16]	7 [0.07]	8 [0.08]	2 [0.02]	8 [0.08]	5 [0.05]	0
Publication type							
Academic journal	78 [0.81]	58 [0.60]	23 [0.24]	15 [0.16]	41 [0.43]	17 [0.18]	6 [0.06]
Evaluation report	3 [0.03]	2 [0.02]	1 [0.01]	1 [0.01]	2 [0.02]	0	0
Dissertation	12 [0.13]	12 [0.13]	2 [0.02]	1 [0.01]	1 [0.01]	6 [0.06]	5 [0.05]
Other	3 [0.03]	2 [0.02]	1 [0.01]	1 [0.01]	2 [0.02]	0	0

NOTES: This table shows the number of studies in our meta-analysis sample that meet the characteristics indicated by the columns and rows. The integer in each cell is the raw count of studies meeting the cell's column and row criteria, while the decimal in brackets is the proportion of studies in our meta-analysis meeting the cell's column and row criteria, the raw count divided by 96, i.e., the total number of studies in our meta-analysis sample.

Table 2
Pooled Tutoring Program Effects on Standardized Test Scores
Categorized by Study Type

Study Type	Pooled Effect Size	Standard Error	Number of Estimates	Number of Studies
All studies	0.37	[0.088]***	732	96
Sample size				
<50	0.45	[0.088]***	134	22
51-100	0.44	[0.076]***	200	24
101-200	0.41	[0.053]***	246	23
201-400	0.30	[0.099]**	89	12
>400	0.25	[0.052]***	63	15
Publication year				
1985-99	0.38	[0.129]**	85	13
2000-09	0.44	[0.055]***	352	39
2010-19	0.32	[0.038]***	295	44
Publication type				
Journal	0.35	[0.034]***	629	78
Non-journal	0.46	[0.089]***	103	18
Bias risk index				
<= 2 of 9 points	0.36	[0.032]***	656	84
> 2 of 9 points	0.43	[0.143]**	76	12

NOTES*:* The table shows pooled effect sizes and associated statistics for subsets of studies falling into the categories listed in the rows. Statistics are generated using single-variable random-effects regressions on effect sizes with inverse propensity weights and robust variance estimation. The bias risk index is a 9-point scale with higher numbers indicating higher risk of bias from measured test scores being more directly related to tutoring material (see paper text for further details). Single, double, and triple asteriks correspond to statistical significance at the 10, 5, and 1 percent level respectively.

Table 3A
Pooled Tutoring Program Effects on Standardized Test Scores
Categorized by Study Type and Tutor

Study Type	Type of Tutor (standard error in square brackets, NS=# of studies, NE=# of estimates)									
	All		Teacher		Paraprof		Nonprof		Parent	
All studies	0.37 [0.032]***	NS=96 NE=732	0.50 [0.075]***	NS=18 NE=179	0.40 [0.040]***	NS=46 NE=360	0.21 [0.064]***	NS=24 NE=127	0.23 [0.114]*	NS=11 NE=67
Subject										
Literacy	0.35 [0.039]***	NS=76 NE=593	0.48 [0.084]***	NS=17 NE=162	0.39 [0.057]***	NS=27 NE=242	0.21 [0.065]***	NS=24 NE=124	0.23 [0.114]*	NS=11 NE=66
Math	0.38 [0.049]***	NS=26 NE=139	0.39 ..	NS=3 NE=17	0.41 [0.055]***	NS=20 NE=118	..		..	
During vs. after school										
During	0.40 [0.037]***	NS=78 NE=617	0.50 [0.079]***	NS=18 NE=179	0.41 [0.041]***	NS=44 NE=349	0.21 [0.088]***	NS=17 NE=89	..	
After	0.21 [0.049]***	NS=17 NE=112	..		0.15 ..	NS=2 NE=11	0.30 ..	NS=6 NE=35	0.16 [0.092]	NS=10 NE=66
Grade level										
Presch-Kind	0.45 [0.051]***	NS=18 NE=116	..		0.41 [0.058]***	NS=10 NE=64	0.46 ..	NS=5 NE=34	0.40 ..	NS=2 NE=15
Grade 1	0.42 [0.041]***	NS=46 NE=374]	0.61 [0.075]***	NS=10 NE=95	0.41 [0.060]***	NS=20 NE=174	0.30 [0.074]***	NS=13 NS=78	0.17 ..	NS=3 NE=27
Grades 2-5	0.29 [0.050]***	NS=50 NE=335	0.40 [0.131]***	NS=9 NE=87	0.41 [0.071]***	NS=20 NE=154	0.13 [0.061]*	NS=17 NE=65	0.21 [0.187]	NS=7] NE=30
Grades 6-11	0.16 ..	NS=8 NE=27	..		0.18 ..	NS=3 NE=13	0.12 ..	NS=3 NE=9	..	

(continued)

Table 3A (continued)
Pooled Tutoring Program Effects on Standardized Test Scores
Categorized by Study Type and Tutor

Study Type	Type of Tutor (standard error in square brackets, NS=# of studies, NE=# of estimates)									
	All		Teacher		Paraprof		Nonprof		Parent	
Tutor to student ratio										
1 to 1	0.38 [0.041]***	NS=67 NE=522	0.59 [0.091]***	NS=13 NE=108	0.46 [0.054]***	NS=22 NE=225	0.21 [0.067]***	NS=23 NE=123	0.23 [0.114]*	NS=11 NE=67
1 to 2	0.29 [0.055]***	NS=14 NE=182	0.56 ..	NS=2 NE=12	0.25 [0.039]***	NS=12 NE=78	..		..	
1 to 3 or more	0.36 [0.063]***	NS=24 NE=182	0.28 ..	NS=4 NE=62	0.38 [0.078]***	NS=19 NE=116	..		..	
Days per week										
1-2 days	0.24 [0.094]***	NS=19 NE=76	..		0.56 ..	NS=5 NE=26	0.14 [0.085]	NS=13 NE=46	..	
3 days	0.34 [0.044]***	NS=33 NE=242	0.81 ..	NS=2 NE=14	0.35 [0.045]***	NS=23 NE=178	0.15 ..	NS=6 NE=32	0.36 ..	NS=2 NE=18
4-5 days	0.41 [0.048]***	NS=46 NE=414	0.49 [0.076]***	NS=15 NE=163	0.38 [0.079]***	NS=19 NE=156	0.60 [0.078]***	NS=6 NE=49	0.19 ..	NS=7 NE=46
Intervention weeks										
20 or fewer	0.39 [0.037]***	NS=69 NE=534	0.55 [0.079]***	NS=15 NE=163	0.41 [0.045]***	NS=40 NE=302	0.14 [0.082]	NS=8 NE=36	0.13 [0.104]	NS=9 NE=53
More than 20	0.29 [0.061]***	NS=27 NE=198	0.30 ..	NS=4 NE=36	0.35 [0.097]***	NS=6 NE=58	0.26 [0.088]***	NS=16 NE=91	0.83 ..	NS=2 NE=14

NOTES: The table shows pooled effect sizes for subsets of studies defined by treatment characteristics. Statistics are generated using single-variable random-effects regressions on effect sizes with inverse propensity weights and robust variance estimation. Unbracketed decimals are effect sizes. Standard errors are in square brackets. Standard error cells for categories yielding fewer than 4 degrees of freedom, as well as categories with too few estimates to run the regression, are marked with ".." Single, double, and triple asteriks correspond to statistical significance at the 10, 5, and 1 percent level respectively.

Table 3B
Pooled Tutoring Program Effects on Standardized Test Scores
Categorized by Study Type and Grade Level

	Grade Level (standard error in square brackets, NS=# of studies, NE=# of estimates)									
Study Type	All		Preschool - Kindergarten		Grade 1		Grades 2-5		Grades 6-11	
All Studies	0.37 [0.032]***	NS=96 NE=732	0.45 [0.051]***	NS=18 NE=116	0.42 [0.041]***	NS=46 NE=374	0.29 [0.050]***	NS=50 NE=335	0.16 ..	NS=7 NE=27
Subject										
Literacy	0.35 [0.039]***	NS=76 NE=593	0.50 [0.059]***	NS=13 NE=92	0.43 [0.049]***	NS=40 NE=335	0.22 [0.049]***	NS=42 NE=256	0.12 ..	NS=5 NE=17
Math	0.38 0.049***	NS=26 NE=139	0.35 ..	NS=5 NE=24	0.38 [0.028]***	NS=9 NE=39	0.44 [0.097]***	NS=14 NE=79	0.20 ..	NS=5 NE=10
During vs. after school										
During	0.40 [0.037]***	NS=78 NE=617	0.48 [0.057]***	NS=14 NE=96	0.45 [0.046]***	NS=38 NE=311	0.32 [0.060]***	NS=39 NE=279	0.09 ..	NS=3 NE=6
After	0.21 [0.049]***	NS=17 NE=112	0.30* ..	NS=4 NE=20	0.28 [0.041]***	NS=8 NE=63	0.19 [0.074]**	NS=10 NE=53	0.29 ..	NS=4 NE=21
Tutor Type										
Teacher	0.50 [0.075]***	NS=18 NE=179	..		0.61 [0.075]***	NS=10 NE=95	0.40 [0.131]***	NS=9 NE=87	..	
Paraprof	0.40 [0.040]***	NS=46 NE=360	0.41 [0.058]***	NS=10 NE=64	0.41 [0.060]***	NS=20 NE=174	0.41 [0.071]***	NS=20 NE=154	0.18 ..	NS=3 NE=9
Nonprof	0.21 [0.064]***	NS=24 NE=127	0.46 ..	NS=5 NE=34	0.30 [0.074]***	NS=13] NE=78	0.13 [0.061]*	NS=17 NE=65	0.12 ..	NS=3 NE=9
Parent	0.23 [0.114]*	NS=11 NE=67	0.40 ..	NS=2 NE=15	0.17 ..	NS=3 NE=27	0.21 [0.187]	NS=7 NE=30	..	

(continued)

Table 3B (continued)
Pooled Tutoring Program Effects on Standardized Test Scores
Categorized by Study Type and Grade Level

	Grade Level (standard error in square brackets, NS=# of studies, NE=# of estimates)									
Study Type	All		Preschool - Kindergarten		Grade 1		Grades 2-5		Grades 6-11	
Tutor to student ratio										
1 to 1	0.38 [0.041]***	NS=67 NE=552	0.51 [0.063]***	NS=12 NE=87	0.44 [0.052]***	NS=35 NE=299	0.25 [0.060]***	NS=35 NE=212	0.13 ..	NS=4 NE=19
1 to 2	0.029 [0.055]***	NS=14 NE=94	0.47 ..	NS=5 NE=36	0.41 ..	NS=4 NE=15	0.25 [0.098]**	NS=8 NE=54	..	
1 to 3 or more	0.36 [0.063]***	NS=23 NE=179	0.32 ..	NS=5 NE=23	0.32 [0.068]***	NS=12 NE=87	0.46 [0.13]***	NS=11 NE=92	..	
Days per week										
1-2 days	0.24 [0.094]**	NS=19 NS=76	0.44 ..	NS=2 NE=10	0.31 ..	NS=5 NE=24	0.24 [0.098]**	NS=18 NS=69	0.14 ..	NS=5 NE=19
3 days	0.34 [0.044]***	NS=33 NE=242	0.40 [0.118]**	NS=7 NE=34	0.34 [0.056]***	NS=21 NE=147	0.37 [0.077]***	NS=16 NE=122	..	
4-5 days	0.41 [0.048]***	NS=46 NE=414	0.49 [0.034]***	NS=10 NE=72	0.48 [0.083]***	NS=21 NE=203	0.28 [0.075]***	NS=16 NE=144	..	
Intervention weeks										
20 or fewer	0.39 [0.037]***	NS=69 NE=534	0.50 [0.059]***	NS=12 NE=78	0.43 [0.051]***	NS=33 NE=256	0.34 [0.059]***	NS=35 NE=257	0.18 ..	NS=2 NE=6
More than 20	0.03 [0.061]***	NS=27 NE=198	0.34 [0.083]**	NS=6 NE=38	0.39 [0.071]***	NS=13 NE=118	0.18 [0.083]***	NS=15 NE=78	0.14 ..	NS=5 NE=21

NOTES: The table shows pooled effect sizes for subsets of studies defined by treatment characteristics. Statistics are generated using single-variable random-effects regressions on effect sizes with inverse propensity weights and robust variance estimation. Unbracketed decimals are effect sizes. Standard errors are in square brackets. Standard error cells for categories yielding fewer than 4 degrees of freedom, as well as categories with too few estimates to run the regression, are marked with ".." Single, double, and triple asteriks correspond to statistical

Table 4
Coefficient Estimates from Regressing Pooled Effect Size on Tutoring Program Characteristics

Independent Variable	Dependent Variable: Pooled Tutoring Program Effect on Standardized Test Scores (Mean=0.37, Std.Dev.=0.424)				
	Model 1	Model 2	Model 3	Model 4	Model 5
Study characteristics					
Ln sample size	-0.06	-0.06	-0.09	-0.08	-0.08
	[0.034]	[0.030]*	[0.040]*	[0.041]*	[0.039]*
Unpublished	0.12	0.09	0.09	0.10	0.13
	[0.102]	[0.100]	[0.107]	[0.111]	[0.121]
High bias risk	-0.07	-0.05	0.03	0.08	0.05
	[0.165]	[0.160]	[0.175]	[0.185]	[0.161]
Dosage					
Ln weeks		0.02	0.06	0.06	0.06
		[0.063]	[0.071]	[0.081]	[0.079]
Days per week		0.07	0.07	0.05	0.02
		[0.034]*	[0.034]*	[0.046]	[0.054]
Tutor type					
Paraprofessional				-0.04	-0.06
				[0.115]	[0.121]
Non-professional				-0.11	-0.17
				[0.149]	[0.169]
Parent				-0.13	-0.19
				[0.295]	[0.301]
Misc intervention characteristics					
Math			0.14	0.11	0.10
			[0.079]*	[0.086]	[0.088]
After School			-0.20	-0.16	-0.15
			[0.091]**	[0.165]	[0.187]
Ratio> 1:1					-0.11
					[0.070]
Presch-Kind					0.09
					[0.071]
Grades 2-5					-0.08
					[0.062]
Grades 6-11					-0.10
					[0.156]
Estimates	732	732	732	732	732
Studies	96	96	96	96	96

NOTES: This table shows coefficents and standard errors (in square brackets) from random effects regressions of effect sizes on intervention and study characteristics. The regressions use inverse propensity weights and robust variance estimation. Single, double, and triple asteriks correspond to statistical significance at the 10, 5, and 1 percent level respectively.

TABLE 5: Large-sample impact evaluations

Study	Intervention	Sample	Tutor type	Subject and level	Dosage	During vs. after school	Tutor-stdnt ratio	Reported findings
Parker et al. (2019)	AmeriCorps Math Tutoring	550 students, 13 schools in Minnesota, 2016-2017 school year	Paraprof (AmeriCorps service fellows)	Math, Grades 4-8	2 45-minute or 3 30-minute sessions per week for ~12 weeks	During	1:2	0.17SD effect on STAR math; increases to 0.24SD under "optimal dosage conditions"
Markovitz et al. (2014)	AmeriCorps (Minnesota Reading Corps)	1,343 students, 23 schools in Minnesota, 2012-2013 school year	Paraprof (AmeriCorps service fellows)	Literacy, Kindergarten - Grade 3	5 20-minute sessions per week for ~16 weeks	During	1:4	1.06 SD for Kindergarten, 0.37 SD for 1st Grade, and 0.10 SD for 3rd Grade; non-signifcant positive coefficient of 0.08 SD for 2nd grade
Jacob et al. (2016)	AmeriCorps (Reading Partners)	1,265 students, 19 schools in 3 states	Nonprof (community volunteers)	Literacy, Grades 2-5	2 45-minutes sessions per weeks for ~30 weeks	Both	1:1	~0.10 SD on SAT-10, TOWRE-2, and AIMSweb oral reading fluency
Lee et al. (2011)	Experience Corps	883 students in 23 schools in Boston, New York City, and Port Arthur (Texas), United States	Nonprof ("older adult" volunteers)	Literacy, 1st grade - 3rd grade	2 to 4 30-40 minute sessions per week for ~16 weeks	During	1:1	0.13SD on Woodcock-Johnson Passage Comrehension; 0.16SD on grade-specific reading skills; and 0.10SD on Woodcock-Johnson Word Attack
Mattera et al. (2018)	High 5s	613 students in 23 New York City schools, 2015-2016 school year	Paraprof, (community members)	Math, Kindergarten	3 30-45 minute sessions per week for ~28 weeks	After	1:4	0.19SD effect on REMA-K; non-signficiant effect of 0.09SD on Woodcock-Johnson Applied Problems
Smith et al. (2013)	Mathematics Recovery	1017 students, 20 schools, 2 unnamed states, 2 school years	Teacher	Math, 1st grade	4-5 30 minute sessions per week for ~12 weeks	During	1:1	0.14SD on a "measure developed by MR" 0.15–0.30SD on independent measures

(continued)

TABLE 5 (CONTINUED)

Study	Intervention	Sample	Tutor type	Subject and level	Dosage	During vs. after school	Tutor-stdnt ratio	Reported findings
Gersten et al. (2015)	Number Rockets	994 students, 76 schools in "4 southwestern and south central states in US", 2008-2009 academic year	Paraprof	Math, 1st grade	3 40 minute sessions per week for ~17 weeks	During	1:2 to 3	0.34SD on TEMA-3
Fuchs et al. (2013)	Number Rockets (Galaxy Math)	591 students, 40 schools in one district, 4 school years	Paraprof	Math, 1st grade	3 30 minute sessions per week for ~16 weeks	During	1:1	Speeded: 0.22-0.87SD; Non-Speeded: 0.19-0.49SD
Sirinides et al. (2018)	Reading Recovery	6,888 students, 1,222 schools across the United States, 2011-2015 school years	Teacher	Literacy, 1st grade	5 30 minute sessions per week for 12-20 weeks	During	1:1	0.48SD effect on ITBS-Total Reading Score
Cook et al. (2015)	Saga Tutoring	2,718 students, 12 schools in Chicago, 2013-2014 school year	Paraprof	Math, 9th grade - 10th grade	5 60-minute sessions per week for 1 school year	During	1:2	0.19-0.31SD effect on math achievement tests; 0.50SD effect on math grades
Miller & Connolly (2013)	Time to Read	734 students, 50 schools in Northern Ireland	Nonprof	Literacy, ages 8-9	1 30-minute session per week for 2 school years	During	1:1	No statistically significant effects
Miller et al. (2012)	Time to Read	512 students, 50 schools in Northern Ireland	Nonprof	Literacy, ages 8-9	2 30-minute sessions per week for 1 school year	During	1:1	No signifcant effects for reading comprehension; 0.14-0.22SD effects on 3 other reading outcomes

FIGURE 1: Analytical model

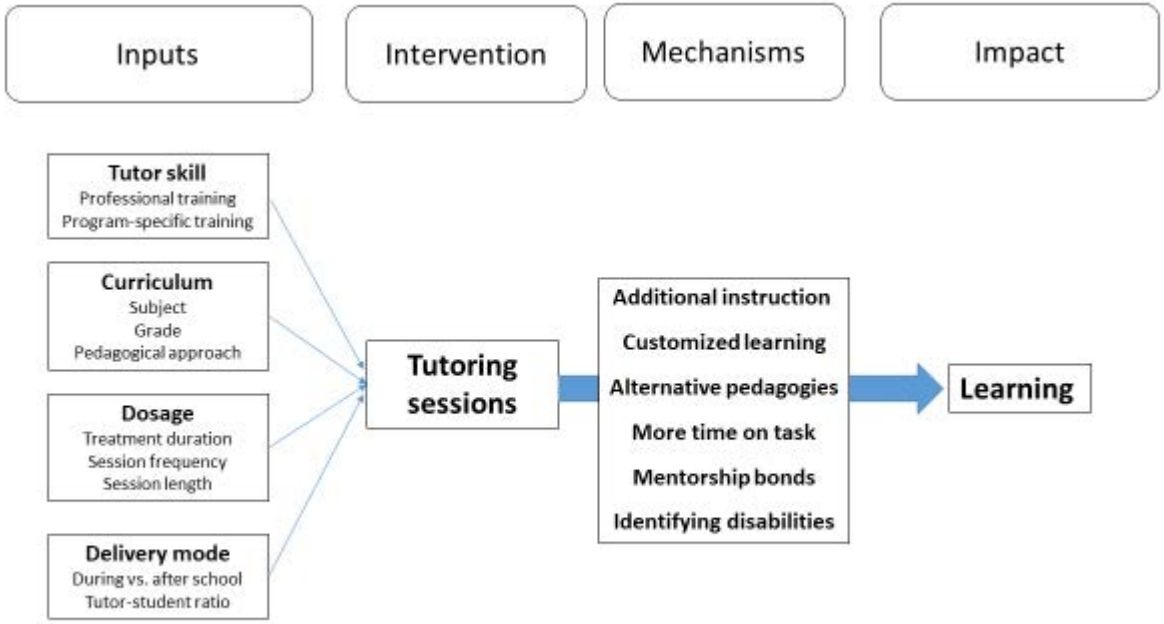
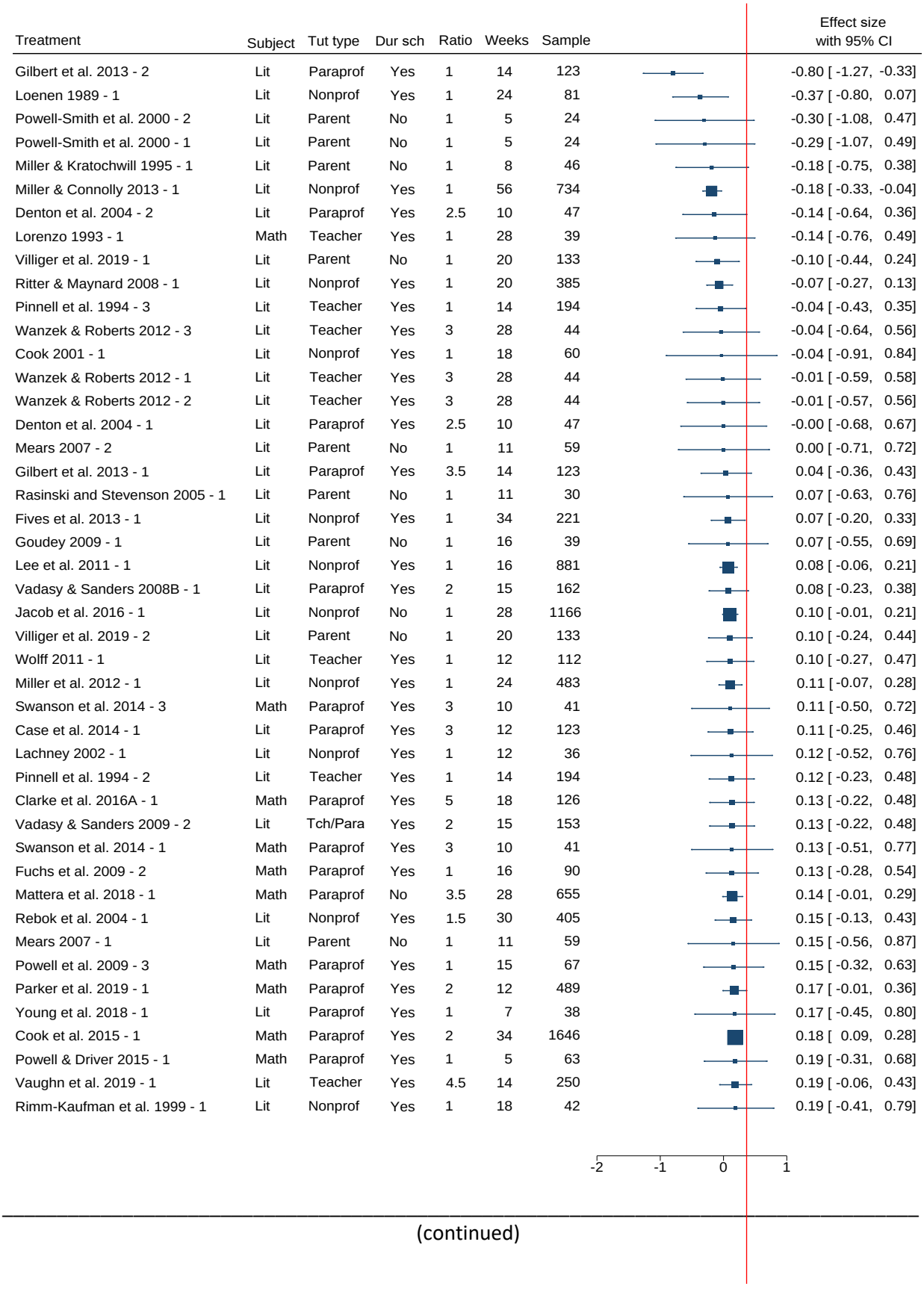
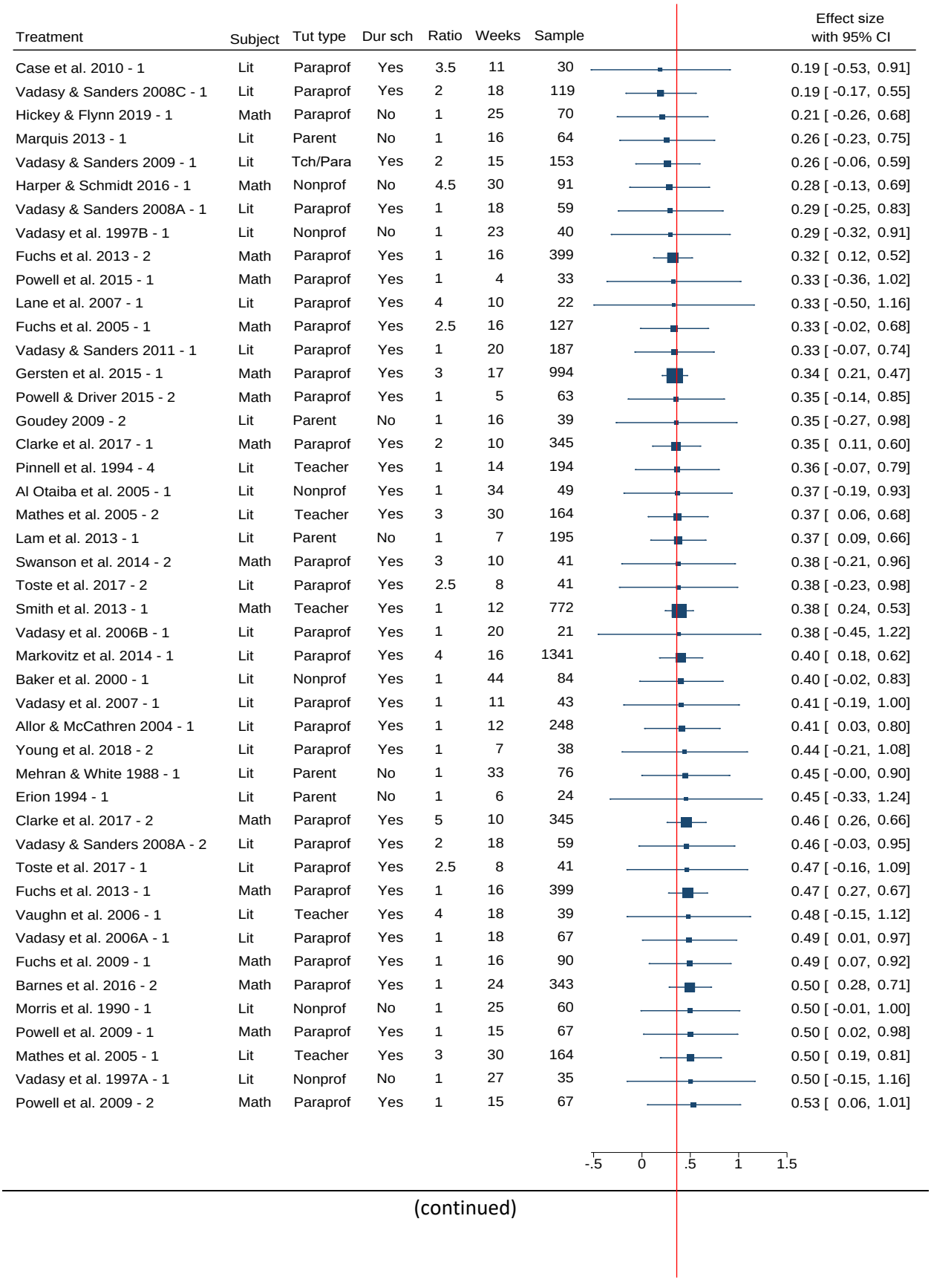


FIGURE 2: Forest plot



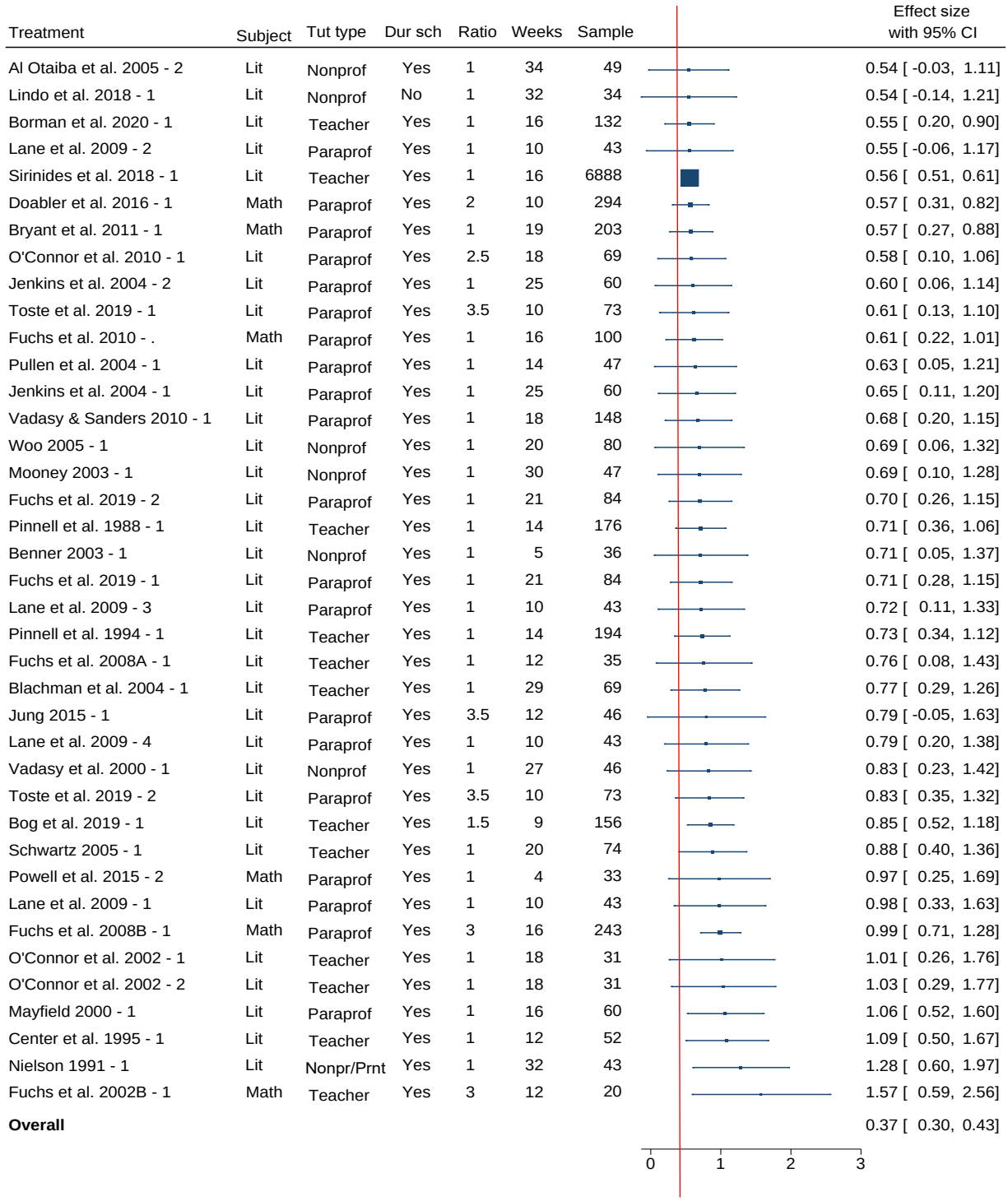
(continued)

FIGURE 2: (CONTINUED)



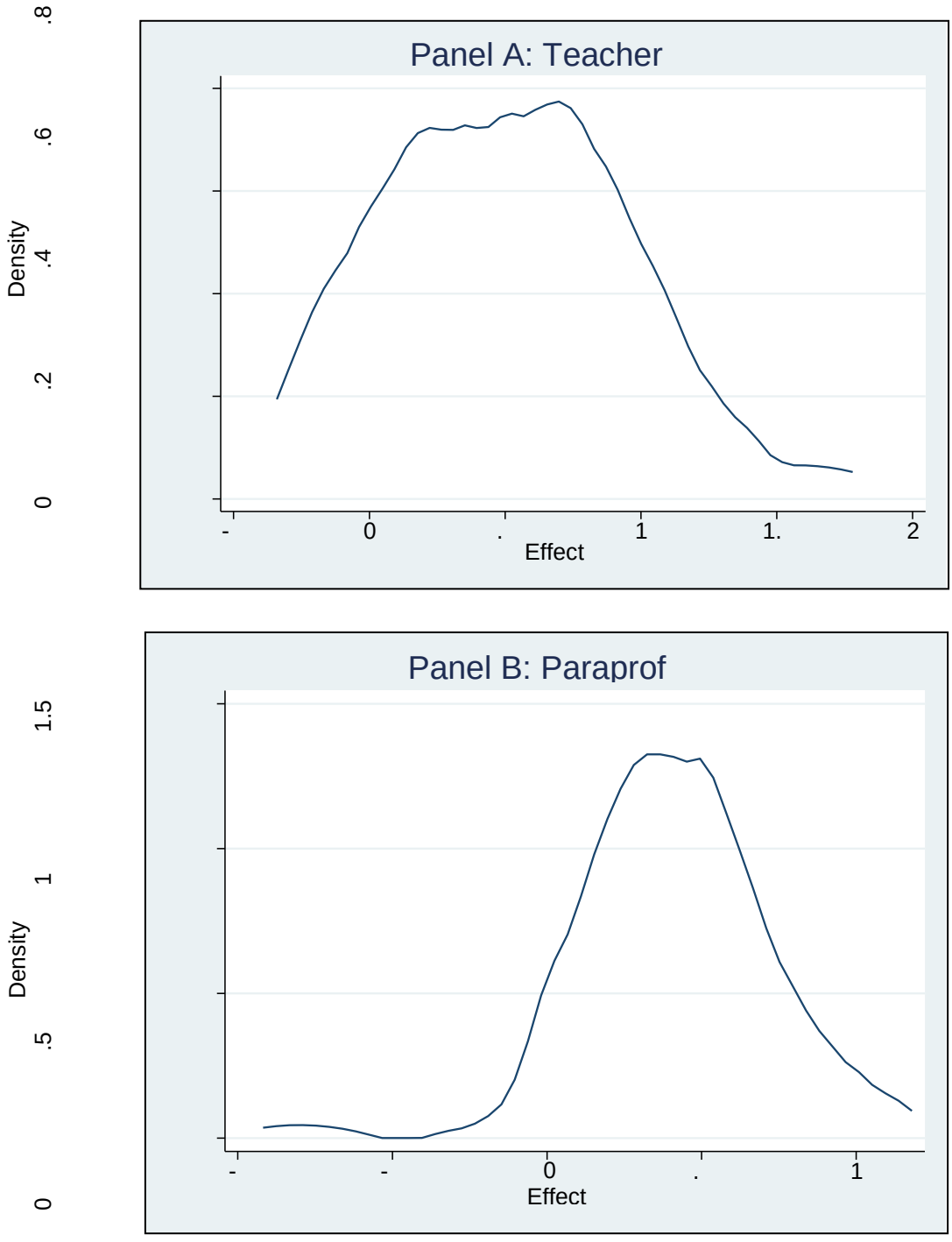
(continued)

FIGURE 2: (CONTINUED)



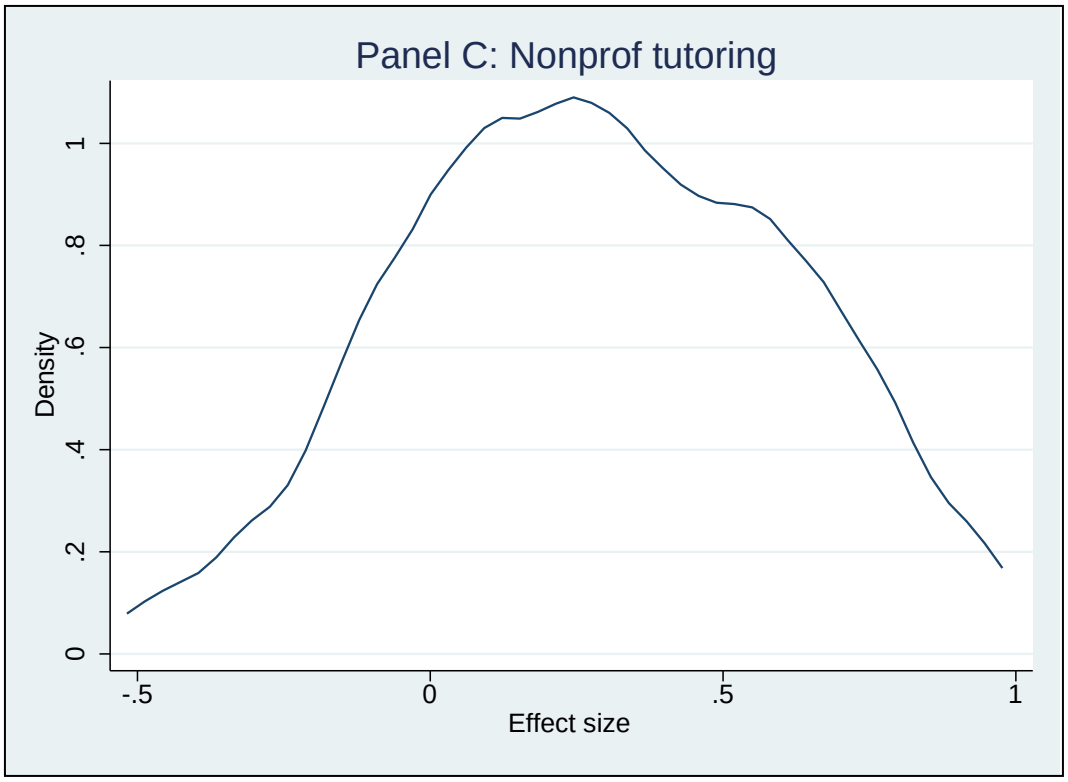
NOTES: This forest plot shows point estimates and 95% confidence intervals for effect sizes of all studies in the meta-analysis sample. Effect sizes and standard deviations are averaged across outcomes within each treatment arm. The overall pooled effect size and confidence interval are taken from Table 2. The vertical red line demarcates this effect.

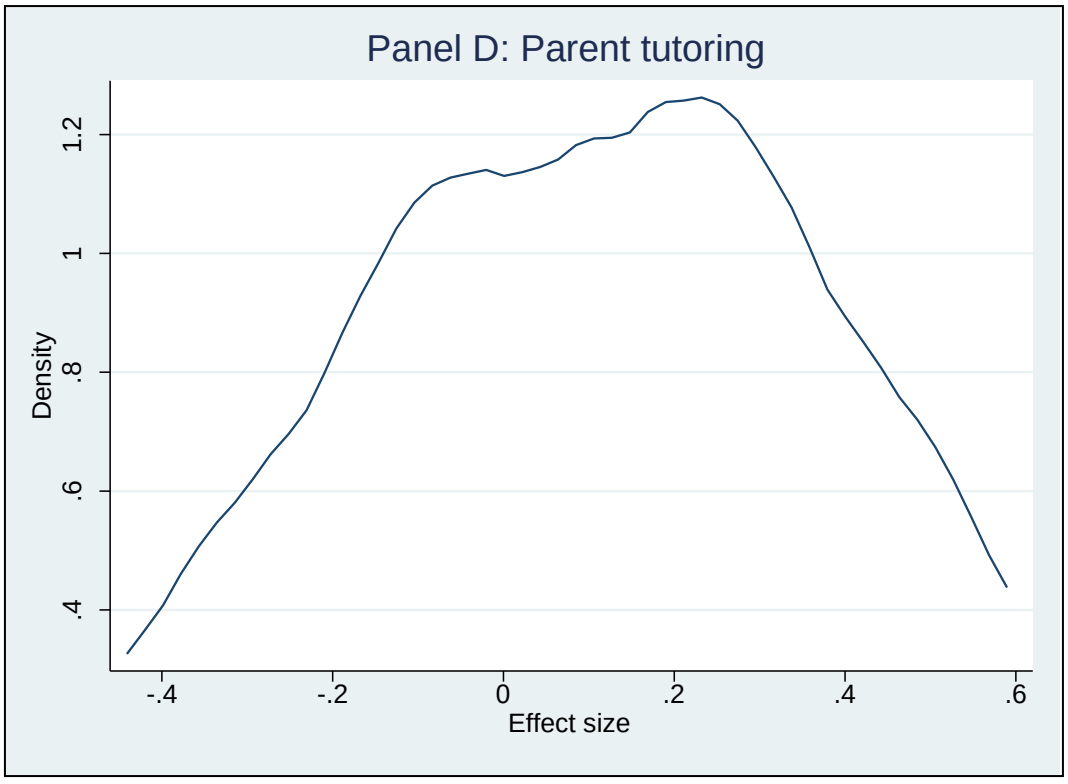
Figure 3
Kernel Density Estimates of Tutoring Program Effects on Standardized Test Scores By Tutoring Type and Subject



(continued)

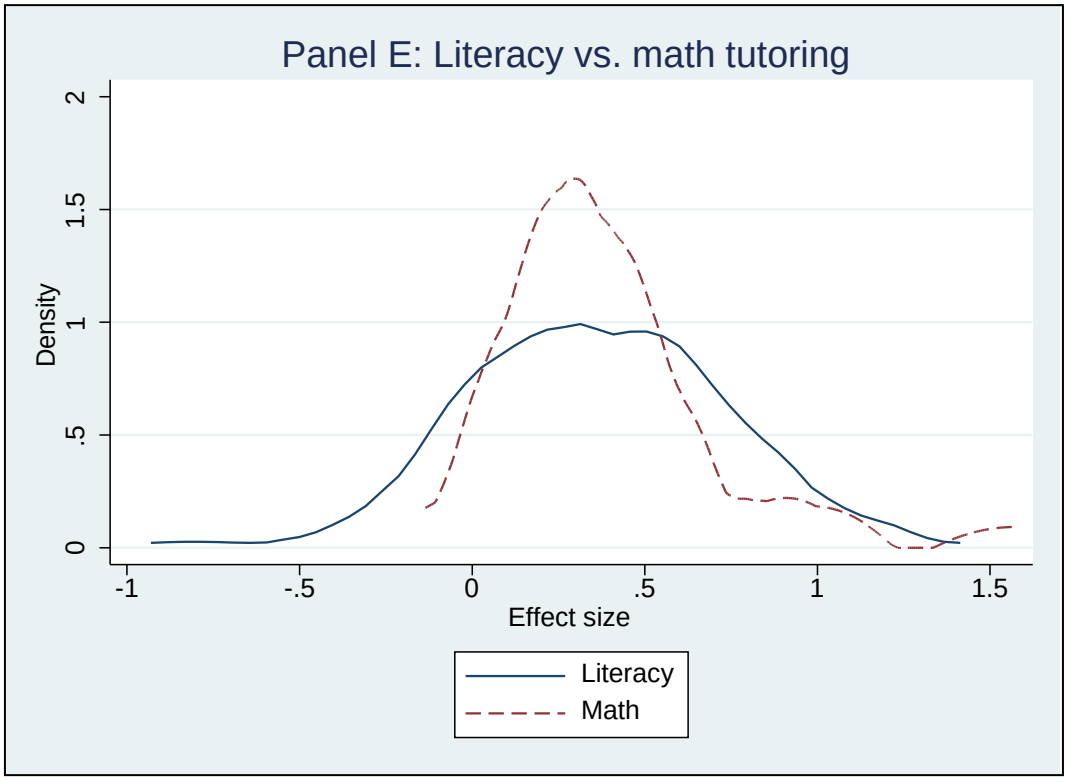
Figure 3 (Continued)





(continued)

Figure 3 (Continued)



NOTES: These kernel density plots depict the distribution of tutoring program effect sizes as they differ across tutor type (Panels A-D) and subject area (Panel E). Effect sizes are unweighted Hedges' *g*, estimated using the procedure described in the methodology section.