Best Evidence
Encyclopedia (BEE)
Empowering Educators with Evidence on Proven Programs
www.bestevidence.org

Effective Programs in Elementary Mathematics:

A Meta-Analysis

Marta Pellegrini

University of Florence

Cynthia Lake

Amanda Neitzel

Robert E. Slavin

Johns Hopkins University

May 2020

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Research and Reform in Education (CRRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Abstract

This article reviews research on the achievement outcomes of elementary mathematics programs. 85 rigorous experimental studies evaluated 64 programs in grades K-5. Programs were organized in 6 categories. Particularly positive outcomes were found for tutoring programs. Positive outcomes were seen in studies focused on professional development for cooperative learning, classroom management, and metacognitive skills. Professional development approaches focused on helping teachers gain in understanding of mathematics content and pedagogy had little impact on student achievement. Professional development intended to help in the adoption of new curricula had a small but significant impact for traditional curricula, but not for digital curricula. Traditional and digital curricula with limited professional development and benchmark assessment programs found few positive effects.

Effective Programs in Elementary Mathematics: A Meta-Analysis

Proficiency in mathematics is essential to success in all quantitative endeavors and occupations, and success in elementary mathematics is of particular importance. In elementary school, students learn basic mathematical ideas and operations, of course, but they also learn that they are either "good at mathematics" or "not good at mathematics." They also learn that mathematics is fun and worthwhile, or that it is tedious and unrewarding. These early learnings can have long term consequences.

According to the National Assessment of Educational Progress (NCES, 2019), fourth grade mathematics scores have improved substantially since 1990, increasing from 13% proficient to 41% in 2019. Most of this gain took place between 1990 and 2009, after which scores have been essentially unchanged. However, U.S. mathematics performance remains mediocre on international comparisons, such as PISA, 2018 (OECD, 2019).

The most serious problem in mathematics education is the continuing inequality in performance between disadvantaged and middle class students. On the 2019 NAEP, only 26% of fourth graders who qualified for free lunches scored at proficient or better, compared to 58% among students who did not qualify for free lunch. These economic differences translate into racial and ethnic group differences. Among fourth graders, 20% of African Americans, 28% of Hispanics, and 24% of Native Americans scored proficient or better, in comparison to 52% of White and 66% of Asian-American students. Disadvantaged and minority students have gained a great deal in mathematics performance since 1990, yet current schooling practices are not enabling them to close gaps with White and Asian-American students. The patterns are similar among eighth graders and beyond, as students who perform poorly in

elementary school lose enthusiasm and confidence in this subject (Ganley & Lubienski, 2016).

The problems of mathematics inequalities and overall performance are longstanding. Yet research in recent years has begun to identify well-defined, replicable elementary mathematics programs that have shown promising outcomes in rigorous experiments (Slavin & Lake, 2008; Jacobse & Harskamp, 2011). A mathematics *program* is a well-specified, replicable set of materials, software, and professional development designed to help teachers improve the mathematics achievement of their students. The findings of program evaluations may validate pragmatic solutions that could have a tangible impact on the mathematics achievement of all students, and especially disadvantaged students. This research has accelerated in amount and quality in recent years.

In light of the importance of elementary mathematics and the social and economic implications of inequalities in mathematics performance by social class and ethnicity, it is clear that substantial investments in improving elementary mathematics performance are needed. Yet which programs and practices are most likely to increase mathematics achievement for all groups?

The importance of evidence for the effectiveness of mathematics programs has increased for U.S. schools as a consequence of the 2015 Every Student Succeeds Act (ESSA). ESSA defines three top levels of evidence for program effectiveness that have important consequences for certain federal funding, especially for low-achieving schools. All three require at least one significant positive effect in well-implemented studies. “Strong” requires at least one randomized study with a positive outcome, “moderate” requires at least one quasi-experimental or matched study with a positive outcome, and “promising” requires at least one correlational

study, with controls for pretests or other covariates, with a positive outcome. ESSA policies make available certain funding for low-achieving schools if they propose to use programs that meet any of these top three ESSA evidence standards, and this has heightened policy interest in rigorous evidence.

Need for This Review

In recent years, several reviews on elementary mathematics programs have been published. Slavin & Lake (2008) identified 87 qualifying studies of outcomes of elementary mathematics programs and concluded that mathematics programs that incorporate cooperative learning, classroom management and motivation, and tutoring had the most positive effects on mathematics achievement. Another review of experimental studies by Jacobse and Harskamp (2011) examined the impact of mathematics interventions in grades K-6 and identified forty studies. The authors reported that small group or individual interventions had greater effects on mathematics achievement than did whole-class programs. Savelsbergh et al. (2016) carried out a meta-analysis of the effect of innovative mathematics interventions on student achievement in grade 1 to 12. Across 19 studies, interventions using technology found moderate positive effects. A review by Gersten et al. (2014) of mathematics professional development found only five studies that met What Works Clearinghouse standards. Higgins, Huscroft-D'Angelo, & Crawford (2019) reported moderate impacts of technology on mathematics achievement.

Although several reviews of elementary mathematics programs have been carried out, the need for high-quality evaluations has particularly increased in recent years. The Institute for Education Sciences (IES), Investing in Innovation (i3) (recently supplanted by a similar program called Education Research and Innovation, or EIR), the National Science Foundation (NSF), and England's Education Endowment Foundation (EEF), have funded numerous

rigorous evaluations of elementary mathematics approaches. Each has highlighted randomized research designs. Other funders, including publishing and software companies, have also supported rigorous research evaluating elementary mathematics programs.

In recent years, there have been major advances in meta-analysis. As the number of rigorous studies of educational innovations has increased, meta-analyses have become more stringent in their inclusion criteria, to focus on the studies capable of making the strongest claims for rigorous evidence linking treatments to outcomes (see Borenstein, Hedges, Higgins, & Rothstein, 2009; Borenstein, Higgins, Hedges, & Rothstein, 2017; Lipsey, 2019; Pigott & Polanin, 2020; Valentine, 2019). In particular, new forms of meta-analysis are using meta-regression to control for moderators, summarize adjusted findings, and make comparisons of effect sizes for various categories of interventions or characteristics of samples, settings, and other moderators. Lynch, Hill, Gonzales, & Pollard (2019) published one of the first meta-analyses of research on STEM subjects to use these methods. They examined research on mathematics and science improvement efforts in grades pre-k to 12. Their analysis of 95 studies focused primarily on characteristics of professional development and curriculum. Their main finding was that professional development specifically used to improve implementation of new curricula had more positive achievement outcomes than professional development to improve teaching in general (unconnected to any particular curriculum), or implementation of curriculum with minimal professional development.

As appropriate to its purposes, the Lynch et al. meta-analysis combined data on many improvement strategies, across all of elementary and secondary mathematics and science. The report focused only on variables coded from the 95 studies that met inclusion criteria, giving little attention to programs (e.g., particular curricula or professional development strategies).

The present meta-analysis uses similar review methodologies, but focuses on programs used to improve mathematics achievement. Evaluating programs, rather than variables alone, can help educators make informed choices of replicable approaches to mathematics improvement. Using meta-regression methods, the present review compares outcomes of categories of programs, as well as moderators (e.g., research design, student grade levels, socio-economic status). Use of meta-regression allows a meta-analysis to discover variables enacted in programs, as well as outcomes of particular programs and categories of programs. The present meta-analysis is neither a critique nor a replication of Lynch et al., but uses the earlier review as a point of departure for a complementary approach to the application of modern meta-analysis and meta-regression, one that emphasizes outcomes of programs and categories of programs in a specific subject and grade span, rather than an emphasis on variables derived from two subjects and all PK-12 grade levels.

Focus of the Review

This review examines research on the effects of elementary mathematics programs on student mathematics achievement. The review considers the strength of evidence supporting particular programs, but it also groups interventions in categories based on their main components to find patterns that may have broader applicability, and it investigates cross-cutting factors and moderators to understand impacts of variables that may contribute to instructional improvement in mathematics.

Method

The present review uses meta-analytic techniques of systematic review and of meta-regression, defined in the following sections.

Inclusion Criteria

The review used rigorous inclusion criteria designed to minimize bias and provide educators and researchers with reliable information on programs' effectiveness. The inclusion criteria are nearly identical to those of Lynch et al. (2019) and of the What Works Clearinghouse (WWC, 2020), with a few exceptions noted below. A PRISMA flow chart (Figure 1) shows the numbers of studies initially found and the numbers winnowed out at each stage of the review. Inclusion criteria were as follows:

1. Studies had to evaluate student mathematics outcomes of programs for elementary schools, grades K-5. Sixth graders were also included if they were in elementary schools. Students who qualified for special education services but attended mainstream mathematics classes were included.
2. Studies had to use experimental methods with random assignment to treatment and control conditions, or quasi-experimental (matched) methods in which treatment assignments were specified in advance. Studies that matched a control group to the treatment group after posttest outcomes were known (post-hoc quasi-experiments or ex post facto designs) were not included.
3. Studies had to compare experimental groups using a given program to control groups using an alternative program already in place, or "business-as-usual".
4. Studies of evaluated programs had to be delivered by ordinary teachers, not by the program developers, researchers, or their graduate students.
5. Studies had to provide pretest data. If the pretest differences between experimental and control groups were greater than 25% of a standard deviation, the study was excluded. Pretest equivalence had to be acceptable both initially and based on pretests for the

final sample, after attrition. Studies with differential attrition between experimental and control groups of more than 15% were excluded.

6. Studies' dependent measures had to be quantitative measures of mathematics performance.
7. Assessments made by program developers or researchers were excluded. This is an important difference from procedures used by Lynch et al. (2019), who included developer/researcher-made measures. The WWC (2020) excludes "over-aligned" measures, but not otherwise acceptable measures made by developers or researchers. The rationale for this exclusion in the current review is that studies have shown that developer/researcher-made measures overstate program outcomes, with about twice the effect sizes of independent measures on average, even within the same studies (Cheung & Slavin, 2016; de Boer et al., 2014). Lynch et al. (2019) included developer/researcher-made measures, but noted that effect sizes were much higher for these measures than for independent measures.
8. Studies had to have a minimum duration of 12 weeks, to establish that effective programs could be replicated over extended periods.
9. Studies could have taken place in North America, Europe, Australia, or New Zealand, but the report had to be available in English. This limitation was intended to focus the review on studies in countries that resemble the U.S. in education systems and cultures. Note that the WWC only includes U.S. studies. In practice, all qualifying studies took place in the U.S., U.K., Canada, The Netherlands, or Germany.
10. Studies had to have been carried out from 1990 through 2019, but for technology a start date of 2000 was used, due to the significant advances in technology since that date.

Literature Search and Selection Procedures

A broad literature search was carried out in an attempt to locate every study that might meet the inclusion requirements. Then studies were screened to determine whether they were eligible for review using a multi-step process that included (a) an electronic database search, (b) a hand search of key peer-reviewed journals, (c) an ancestral search of recent meta-analyses, (d) a web-based search of educational research sites and educational publishers' sites, and (e) a final review of citations found in relevant documents retrieved from the first search wave.

First, electronic searches were conducted in educational databases (JSTOR, ERIC, EBSCO, PsycINFO, ProQuest Dissertations & Theses Global) using different combinations of key words (e.g., "elementary students," "mathematics", "achievement", "effectiveness", "RCT", "QED"). We also searched in recent tables of contents of eight key mathematics and general educational journals from 2013 to 2019: *American Educational Research Journal, Educational Research Review, Elementary School Journal, Journal of Educational Psychology, Journal of Research on Educational Effectiveness, Journal for Research in Mathematics Education, Learning and Instruction,* and *Review of Educational Research.* We investigated citations from previous reviews of elementary mathematics programs (e.g., Dietrichson, Bøg, Filges, & Klint Jørgensen, 2017; Gersten et al., 2014; Jacobse & Harskamp, 2011; Li & Ma, 2010; Lynch et al., 2019; Savelsbergh et al., 2016).

We were particularly careful to be sure we found unpublished as well as published studies, because of the known effects of publication bias in research reviews (Cheung & Slavin, 2016; Chow & Ekholm, 2018; Polanin, Tanner-Smith, & Hennessy, 2016). We

searched for studies published online by funding agencies such as i3, IES, NSF, and EEF, and for studies reviewed by the What Works Clearinghouse (WWC) and Evidence for ESSA (www.evidenceforessa.org). We also visited the websites of educational societies (American Educational Research Association and Society for Research on Educational Effectiveness) to search for conference presentations. Finally, we reviewed citations of documents retrieved from the first wave to search for any other studies of interest.

A first screen of each study was carried out by examining the title and abstract using inclusion criteria. Studies that could not be eliminated in the screening phase were located and the full text was read by one of the study authors. We further examined the studies that were believed to meet the inclusion criteria and those where inclusion was possible but not clear. All of these studies were examined by a second author to determine whether they met the inclusion criteria. When the two authors were in disagreement, the inclusion or exclusion of the study was discussed with a third author until consensus was reached.

After removing duplicate studies, these search strategies yielded 18,642 studies for screening. The screening phase eliminated 13,366 studies, leaving 1,120 full-text articles to be assessed for eligibility. Of these full-text articles that were reviewed, 1,038 studies did not meet the inclusion criteria, leaving 82 contributions included in this review, with two studies including multiples interventions for a total number of 85 studies (see Figure 1).

Coding

Studies that met the inclusion criteria were coded by one of the authors of the review. Then codes were verified by another author. As for the inclusion of the studies, disagreements were discussed with a third author until consensus was reached.

Data coded included: program components, publication status, year of publication, study design, study duration, sample size, grade level, participant characteristics, outcome measures, and effect sizes.

We also identified variables that could possibly moderate the effects in the review distinguishing between substantive factors and methodological factors. Substantive factors are related to the intervention and the population characteristics. The factors coded were grade level (K-2 vs. 3-6), student achievement levels (low achievers vs. average/high achievers), socio-economic status (low SES vs. moderate/high SES), and study locations in the U.S. vs. other countries. Methodological factors included research design (quasi-experiments vs. randomized studies). For tutoring programs we also coded the group size (one-to-one vs. one-to-small group) and the type of provider (teacher, teaching assistant, paid volunteer, or unpaid volunteer).

Effect Size Calculations and Statistical Procedures

Effect sizes were computed as the mean difference between the posttest scores for individual students in the experimental and control groups after adjustment for pretests and other covariates, divided by the unadjusted standard deviation of the control group's posttest scores. Procedures described by Lipsey and Wilson (2001) were used to estimate effect sizes when unadjusted standard deviations were not available.

Statistical significance is reported for each study using procedures from the What Works Clearinghouse (WWC, 2020). If assignment to the treatment and control groups was at

the individual student level, statistical significance was determined by using analysis of covariance (ANCOVA), controlling for pretests and other factors. If assignment to the treatment and control groups was at the cluster level (e.g., classes or schools), statistical significance was determined by using multilevel modeling such as Hierarchical Linear Modeling (HLM, Raudenbush & Bryk, 2002). Studies with cluster assignments that did not use HLM or other multi-level modeling but used student-level analysis were re-analyzed to estimate significance with a formula provided by the WWC (2020) to account for clusters.

Mean effect sizes across studies were calculated after assigning each study a weight based on inverse variance (Lipsey & Wilson, 2001), with adjustments for clustered designs suggested by Hedges (2007). In combining across studies and in moderator analysis, we used random-effects models as recommended by Borenstein et al. (2009).

Meta-regression

We used a multivariate meta-regression model with robust variance estimation (RVE) to conduct the meta-analysis (Hedges et al., 2010). This approach has several advantages. First, our data included multiple effect sizes per study, and robust variance estimation accounts for this dependence without requiring knowledge of the covariance structure (Hedges et al., 2010). Second, this approach allows for moderators to be added to the meta-regression model and calculates the statistical significance of each moderator in explaining variation in the effect sizes (Hedges et al., 2010). Tipton (2015) expanded this approach by adding a small-sample correction that prevents inflated Type I errors when the number of studies included in the meta-analysis is small or when the covariates are imbalanced. We estimated three meta-regression models. First, we estimated a null model to produce the average effect size without adjusting for any covariates. Second, we estimated a meta-regression model with the identified moderators of

interest and covariates. Third, we estimated an exploratory meta-regression model which added tutoring provider as a moderator. Due to the small sample size, this model is considered exploratory and results of statistical tests such as p-values are not reported. All moderators and covariates were grand-mean centered to facilitate interpretation of the intercept. All reported mean effect sizes come from this meta-regression model, which adjusts for potential moderators and covariates. The packages *metafor* (Viechtbauer, 2010) and *clubSandwich* (Pustejovsky, 2020) were used to estimate all random-effects models with RVE in the R statistical software (R Core Team, 2020).

Categories of Mathematics Programs

Studies that met the inclusion criteria were divided into categories according to the main and most distinctive components of the programs. Categories were formed based on four widely-held (and not mutually-exclusive) explanations for the principal problems intended to be solved in elementary mathematics improvement strategies: Low student achievement, need to enhance teacher understanding of mathematics content and pedagogy, need to enhance teacher understanding and use of methods to increase student motivation, engagement, and cognition, and need for teachers to use standards-based curricula. These are shown in Figure 2.

Category assignments were based on independent readings of articles and websites by the authors. All authors read all accepted studies, and if there were disagreements about categorizations they were debated and determined by consensus among all authors. The categories and their theoretical rationales were as follows.

1. Tutoring. Tutoring refers to one-to-one or one-to-small group instruction intended to help students struggling in mathematics. Tutoring may involve one teacher or one teaching

assistant (paraprofessional) with one student, or one teacher or teaching assistant with a very small group of students, usually from two to six at a time.

There are several ways in which tutoring is likely to improve student mathematics outcomes. First, tutoring (especially one-to-one) permits tutors to completely adapt their instruction to the needs of the student(s). Well-trained tutors are able to start with struggling students where they are and move them forward rapidly, instead of leaving them to flounder in the regular class with challenges too far above their current levels of prior knowledge. Second, tutors are likely to be able to build close personal relationships with the tutored student(s), giving them attention and praise that many students crave.

Tutoring was not included in the review by Lynch et al., who focused only on programs delivered to entire classes. However, previous reviews of research on elementary mathematics approaches have found that tutoring is among the most effective (e.g., Slavin & Lake, 2008; Jacobse & Harskamp, 2011). Tutoring has also been found to be very effective in elementary reading (e.g., Neitzel, Lake, Pellegrini, & Slavin, 2020; Wanzek et al., 2016).

2. Professional Development Focused on Mathematics Content and Pedagogy provides intensive content-focused professional development (PD) intended to advance teachers' understanding of current standards-based content and effective pedagogy. The theory of action emphasizes giving teachers knowledge about mathematics content and about ways of explaining it, rather than new texts or new software alone (Cohen & Hill, 2000; Desimone & Garet, 2015).

3. Professional Development Focused on Classroom Management, Motivation, and Cognition. This diverse category includes programs that provide teachers with professional development and materials to help them implement innovations in classroom organization and

management, such as cooperative learning (e.g., Slavin, 2017), classwide behavior approaches (e.g., Weis et al., 2015), and strategies designed to improve students' meta-cognitive capacities (e.g., Muijs & Bokhove, 2020; Roy et al., 2019).

4. Professional Development Focused on Implementation of Traditional and Digital Curricula provides teachers with moderate to extensive professional development (at least two days, or 15 hours, combining training and follow-up coaching) to support informed, thoughtful implementation of innovative traditional or digital curricula. There were two subcategories: a) Professional Development Focused on Implementation of Traditional Curricula (with minimal use of technology), and b) Professional Development Focused on Implementation of Digital Curricula, such as computer-assisted instruction. The first of these corresponds to the most effective type of approaches across all of STEM identified by Lynch et al. (2019).

5. Traditional and Digital Curricula With Limited Professional Development includes two subcategories: a) Traditional curricula (textbooks with associated teaching materials), and b) Digital curricula. Limited professional development (less than two days or 15 hours) was typically included in such strategies (if extensive professional development had been provided, programs would have been included in Category 4).

6. Benchmark Assessments consist of tests given periodically (three to five times a year) to find out how students are proceeding toward success on state standards. The rationale is to give teachers and school leaders early information on student performance so they can make changes well before state testing (e.g., Konstantopolous, Miller, van der Ploeg, & Li, 2016).

Results

A total of 85 studies evaluating 64 programs met the inclusion standards of this review. The studies included were of high methodological quality: 72 (85%) of the studies were randomized trials and 13 (15%) were quasi-experimental studies. 73 (86%) of the studies were reported in 2010 or later, indicating the extraordinary pace at which rigorous studies of elementary mathematics are appearing. Table 1 shows the meta-regression outcomes. The full model controlled for program category and subcategory, research design, grade level, student achievement level, SES, U.S. vs. other countries, and tutoring group size. Table 2 shows adjusted means for each category and subcategory. Tables 3 to 8 summarize the main characteristics and outcomes of the individual studies, grouping them by category, and Table 9 shows effects of moderators. Across all included studies of programs on elementary mathematics, we found an average weighted effect size of +0.09, $p < .01$ ($k = 85$), with outcomes that vary substantially among different categories.

Tutoring Programs

Twenty-three studies evaluated tutoring programs. Combining all forms of tutoring, the mean effect size was +0.20, $p < .01$ ($k = 21$). Table 3 shows the tutoring programs, study details, and findings. Eight of these evaluated face-to-face, one-to-one tutoring. An additional study evaluated one-to-one tutoring from tutors in India or Sri Lanka delivered online to students in the U.K., and another evaluated cross-age peer tutoring. These two approaches were so different from other tutoring models and had such limited evidence (one study each) that they are not averaged with the others. Thirteen studies evaluated programs taught by tutors to small groups. Overall, the weighted mean effect size for one-to-one face-to-face tutoring was +0.19, $p < .01$ ($k = 8$), while the single study of one-to-one online tutoring program had an effect size of -0.03 and the one study of cross-age peer tutoring had an effect

size of +0.02. One-to-one tutoring by certified teachers (*ES* = +0.22) (*k* = 2), and by teaching assistants (*ES* = +0.18) (*k* = 5) were not significantly different from each other in outcomes in the exploratory model. Teaching assistants were relatively well qualified (e.g., most had bachelor's degrees), and both certified teachers and teaching assistants used structured programs and received extensive professional development. One program used paid AmeriCorps volunteers[1] as tutors, and the ES was +0.20.

Tutoring to small groups had an overall mean effect size of +0.30, $p < .01$ (*k* = 13). Surprisingly, outcomes of one-to-small group tutoring using structured programs were (non-significantly) higher than those of one-to-one tutoring. The only one-to-small group program that used certified teachers (*ES* = +0.34, *k* = 1) was similar in outcomes to one-to-small group approaches that used teaching assistants as tutors (*ES* = +0.30, $p < .01$, *k* = 12). The numbers of studies in some categories of tutoring were small, so these findings must be interpreted with caution, but it is interesting that while all forms of face-to-face tutoring by paid adults had quite positive impacts on achievement, the outcomes were highest for one-to-small group approaches.

Professional Development Focused on Mathematics Content and Pedagogy

Nine studies evaluated nine programs focused on teacher professional development to improve teachers' knowledge of mathematics content and content-specific pedagogy. The programs use various types of support for teachers such as workshops, training, continuous professional development, in-school support, and coaching. They may focus on improving

[1] AmeriCorps is a U.S. program that recruits and trains volunteers to provide services (such as tutoring) to their communities. Volunteers receive stipends and other benefits.

teachers' content knowledge, content-specific pedagogy, general pedagogy, or some combination of these. Table 4 shows the programs, study details, and outcomes. The adjusted mean effect size was +0.03, n.s. (k = 9) for all professional development programs focused on mathematics content and pedagogy.

Professional Development Focused on Classroom Management, Motivation, and Cognition

Professional development approaches in this category focused on helping teachers use models such as cooperative learning, classroom management, and teaching focused on meta-cognitive skills and working memory in elementary mathematics (see Table 5). Across eight studies of seven diverse programs, the average effect size for mathematics was +0.19, $p < .01$ (k = 8).

Professional Development Focused on Implementation of Traditional and Digital Curricula

12 studies evaluated 10 programs in which significant professional development supported the implementation of new curricula or software. Table 6 shows study details and outcomes. The mean effect size was +0.02, $p < .01$ (k = 12). Effect sizes averaged +0.12, $p < .01$ (k = 7) for traditional curricula, but +0.01, n.s. (k = 5) for digital curricula.

Traditional and Digital Curricula With Limited Professional Development

Twenty-nine studies evaluated 18 mathematics curricula, primarily traditional or digital textbooks with teacher materials and limited professional development. Study details and outcomes are summarized in Table 7. Across all qualifying studies, the adjusted mean effect size was +0.03, n.s. (k = 29). Fifteen studies of traditional curricula, mostly textbooks, found a

mean effect size of +0.03, n.s. (k = 15), and 14 studies that evaluated digital curricula found a mean effect size of +0.07, p < .05. (k = 14).

Benchmark Assessments

Four studies evaluated three programs that use benchmark assessments, summarized in Table 8. The studies found a mean effect size of 0.00, n.s. (k = 4).

Moderator Analyses

Random-effects models were used to carry out moderator analyses, which identify substantive and methodological factors that contribute to positive outcomes (see Table 9). Moderator analyses including all studies were conducted. An exploratory model was used to examine the effect of tutoring provider, by adding it to all other identified moderators.

Research design. As reported in previous studies, effect sizes may vary according to research design. Cheung & Slavin (2016) and de Boer et al. (2014) found that quasi-experiments across all subjects and grade levels, PK-12, produce a significantly higher effect size than randomized studies, on average, although others, such as Lipsey & Wilson (2001), have not found this difference. Differences in effect sizes between studies that used randomized designs (ES = +0.08, p < .01, k = 72) and studies that used quasi-experimental designs incorporating matching (ES = +0.18, p < .01, k = 13) were tested. This difference (β = 0.10) was significant (p < .05).

Grade levels. To determine if different grade levels may be a source of variation, we divided the studies into those that took place in K to 2 or in 3 to 6. The mean effect size for K-2 outcomes (ES = +0.09, p < .01, n = 68) was very similar to the mean effect size for 3-6 outcomes (ES = +0.10, p < .01, n = 82). When compared to outcomes including K-6 (ES = +0.10, p < .01, n = 27), neither were significantly different.

Student achievement level. Outcomes including all students had a mean effect size of +0.08, $p < .01$ ($n = 114$). This was not significantly different from either outcomes for low achievers ($ES = +0.14$, $p < .01$, $n = 48$) or outcomes for moderate and high achievers ($ES = +0.07$, $p < .05$, $n = 15$).

Socio-economic status (SES). Study samples were defined as low-SES if the proportion of students receiving free or reduced-priced meals was at or above the 75th percentile of school rates of free- or reduced- price meals participation at the national level (76% for the U.S., 21% for England). Mean effect sizes for outcomes of mixed SES populations were +0.09, $p < .01$ ($n = 54$). The mean effect size for low SES students was +0.08, $p < .05$ ($n = 53$), and for moderate/high SES students it was +0.11, $p < .01$ ($n = 76$). The differences between mixed and low SES students ($\beta = -0.01$, n.s.) and mixed and moderate/high SES students ($\beta = 0.02$, n.s.) were not statistically significant.

U.S. vs. Other Countries. Of the 85 qualifying studies, 63 took place in the U.S., 19 in England, one in the Netherlands, one in Germany, and one in Canada. Mean effect sizes were nearly identical for U.S. and non-U.S. studies: +0.10, $p < .01$ for U.S ($k = 63$), +0.07, $p < .01$ for non-U.S. ($k = 22$). This difference ($\beta = -0.03$, n.s.) was not statistically significant.

Tutoring-Specific Moderators

Tutoring group size. The impacts of tutoring provided in a one-to-one format ($ES = +0.19$, $p < .01$, $k = 8$) were compared to those for tutoring provided in small-group settings ($ES = +0.30$, $p < .01$, $k = 13$). Outcomes were not significantly different ($\beta = 0.11$, n.s.).

Tutoring provider. Because there were small numbers of studies of tutoring with different providers, this moderator was explored in a separate exploratory model still containing all other moderators and covariates. The mean effect sizes for five different

combinations of providers and group size (one or small group) are shown in Table 9 as an exploratory analysis, and statistical tests such as p values are not reported.

Among the tutoring studies, the outcomes of tutoring provided by teachers (*ES* = +0.23, *k* = 3) was similar to those of tutoring provided by teaching assistants (*ES* = +0.19, *k* = 17).

Discussion

This review of evaluations of elementary mathematics programs found 85 studies of very high methodological quality. The studies were mostly randomized and large-scale, increasing the likelihood that their findings will replicate in large-scale applications in practice. Collectively, the studies found that it matters a great deal which programs and which types of programs elementary schools use to teach mathematics, especially for low-achieving students.

The findings of the current study provide some support for the conclusions of Lynch et al. (2019). Of course, the present study focused only on elementary mathematics, and Lynch et al. addressed science as well as mathematics in grades pre-K to 12, so this is not a head-to-head comparison. But the relative outcomes are nevertheless interesting.

Both Lynch et al. and the present study found small, non-significant impacts for professional development services without a strong link to new curriculum, and both found small, non-significant impacts of implementation of traditional or digital curricula with a limited focus on professional development (less than 15 hours). Lynch et al. found positive effects for strategies that focused professional development on the implementation of new curricula. The present study also found small but significant positive effects of strategies that devote extensive professional development to adoption of traditional curricula (*ES* = +0.12, $p < .01$), but found an effect size near zero for programs that provide extensive professional development to support use of digital curricula. The present meta-

analysis also found significant positive effects of professional development to help teachers improve classroom management, motivation, and cognition (*ES* = +0.19, $p < .01$). Forms of cooperative learning were most common among such studies. The Lynch et al. (2019) meta-analysis did not identify a category of professional development focusing on motivation or cognition, because its focus was on the interaction of professional development and curriculum.

The other category of approaches that had the largest and most robust impacts was tutoring, excluded from the Lynch et al. review because of its focus on whole-class instruction. One-to-one tutoring by face-to-face adult tutors and one-to-small group tutoring were particularly effective. It was interesting to find that the effect size for one-to-small group tutoring (*ES* = +0.30, $p < .01$, $k = 13$) was larger than that for one-to-one (*ES* = +0.19, $p < .01$, $k = 8$). Teachers (*ES* = +0.23) and teaching assistants (*ES* = +0.19) appear equally effective as tutors, on average, but this result should be interpreted cautiously due to the exploratory nature of that analysis. In contrast, on-line tutors and cross-age peer tutors did not show promising impacts. The findings suggesting that the least expensive tutoring format, one-to-small group tutoring by teaching assistants, was quite effective (*ES* = +0.30, $k = 12$) suggests that tutoring (by teaching assistants to small groups) could be a very cost-effective service for students struggling in mathematics, and could therefore be practicably offered to larger numbers of students than has previously been thought possible.

Theorists have long assumed that tutoring works well because the tutor can fully adapt to the learning needs of students (e.g., Wanzek et al., 2016). Yet effect sizes for all studies using digital curricula had effect sizes near zero (see Tables 6 and 7). Most technology use in mathematics adapts the level and pace of instruction to the needs of each student, as does tutoring, yet adaptive technology does not have notably large impacts. This difference in

outcomes for two adaptive solutions calls into question the explanation of tutoring effects as being primarily due to adaptation to individual needs.

If tutoring does not mainly owe its substantial effects to its adaptation to student needs, then why does it work? One additional explanation may be that tutoring provides struggling students with individual attention from caring tutors. Educational technology can be fun and engaging, but computers cannot form significant relationships with children. The pattern of findings and the contrasts in outcomes among seemingly similar interventions support a conclusion that, assuming well-trained tutors and well-structured materials, adult-student relationships are essential to the unique success of tutoring. Perhaps these relationships are especially important for struggling students, who may receive less positive attention in class than others. In any case, further research on tutoring is clearly needed to understand what seems such a simple question: Why does tutoring work?

The discrepancy in outcomes was striking between studies of professional development focused on building teachers' knowledge of mathematics content and pedagogy and those of professional development focused on helping teachers implement innovations in classroom organization and management. One extraordinary example is a study of Intel Math (Garet et al., 2016), which provided 93 hours of in-service to teachers of grades K-8 to improve their understanding of mathematics content and pedagogy. A one-year cluster randomized evaluation with 165 teachers found small but significantly *negative* impacts on state tests (ES = -0.06, $p < .05$), and nearly identical but non-significant negative effects on NWEA Mathematics. Several studies found significant positive impacts on teachers' knowledge of mathematics, but this did not transfer to improvement in student achievement. Not one of the 9 studies of professional development methods focused on mathematics

content and pedagogy achieved statistical significance, and the mean was only +0.02. It is of course important for teachers to know and apply appropriate mathematics content and content-specific pedagogy, but perhaps this is not enough if the student experience is not fundamentally changed. Another possibility is that teachers in the control groups already knew a great deal about mathematics content and pedagogy, so further professional development in these areas may not make much difference. Clearly, a deeper look into programs of this kind is warranted.

Studies of traditional and digital mathematics curricula with limited professional development found very small impacts (mean *ES* = +0.03, *k* = 29, n.s.). Most of the mathematics curriculum studies just compared a new textbook or digital curriculum (and associated add-ons) to existing textbooks or software, so it is not surprising to see few differences in outcomes. Similarly, studies of benchmark assessments found a near-zero mean effect size of 0.00 (*k* = 4, n.s.).

One interesting finding from the present review relates to technology in mathematics education, which has been reviewed previously by Cheung & Slavin (2013); Higgins et al. (2019); Li & Ma, 2010; and Savelsbergh et al. (2016). Technology is now used in so many ways that it no longer makes sense to make generalizations about what technology can or cannot accomplish in mathematics education. Nevertheless, it is striking how weak the evidence base for technology is. The present research adds to the evidence on technology applications in several ways. First, the category of Professional Development Focused on the Implementation of Traditional and Digital Curricula had two subcategories, identifying programs with or without an emphasis on technology (see Table 6). Programs that provided extensive professional development to support traditional curricula, essentially textbooks, had a modest positive impact

on mathematics achievement, averaging +0.12. However, professional development supporting programs with a strong focus on technology had an average effect size of +0.01. Further, among programs with limited professional development, both traditional curricula (*ES* = +0.03) and digital curricula (*ES* = +0.07) had minimal effect sizes (see Table 7). Between these two categories, there were 19 technology-focused programs and 22 traditional ones, meaning that there were enough studies of each kind for adequate statistical power. Yet programs emphasizing technology showed no advantage in outcomes compared to programs without a technology focus. Especially in mathematics, which seems to lend itself to technology more than any other subject, to find so little evidence supporting the value-added of technology is disturbing. Perhaps approaches that use technology will be created and successfully evaluated in the future, but the evidence of the present review would suggest that the ways technology is currently being applied in mathematics are not making much of a difference in outcomes.

As noted previously, the most important problem in U.S. mathematics education is the inequality between advantaged and disadvantaged students. The evidence from the present review suggests approaches with the strongest impacts for low achievers were one-to-one and one-to-small group tutoring. These very effective strategies are only used with low achievers, as a practical matter. The positive outcomes for small group tutoring by teaching assistants suggest that tutoring may be an economically feasible way to increase low achievers' mathematics achievement. Across all approaches, effects were larger for low achievers (*ES* = +0.14) than for others (moderate/high achievers: *ES* = +0.07, mixed achievers: *ES* = +0.08), suggesting additional pragmatic methods of increasing means while narrowing gaps.

If this pattern of findings is replicated in future research, it would suggest that in addition to tutoring for students struggling in mathematics, professional development in

strategies focused on motivation, engagement, and metacognitive skills should be a focus of mathematics improvement, especially for low-achieving students.

Conclusion

This meta-analysis provides encouraging findings, suggesting that low achievers can make substantial gains in mathematics if they receive relatively cost-effective small group tutoring. Promising outcomes were also achieved by programs that emphasize cooperative learning, classroom management, and teaching of metacognitive skills. These findings support a belief that long-standing inequalities in mathematics achievement can be overcome using proven, replicable strategies and by professional development focused on implementation of traditional curricula.

Limitations

This review is focused on rigorous experimental studies evaluating student mathematics outcomes. Although other research designs, such as qualitative and correlational research, can add depth and understanding of the effects of mathematics programs, for policy purposes it is crucial to evaluate programs in comparison to control groups receiving traditional teaching, according to their impacts on quantitative measures in rigorous designs. In addition, the review excludes measures made by researchers or developers of the programs. These measures may be of theoretical interest, but are often unfair to control groups because they are likely to be aligned with the content taught in the experimental but not in the control group.

References

**Studies included in the meta-analysis.*

*Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders* (NCEE 2011-4001). Washington, DC: U.S. Department of Education.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

*Boylan, M., Demack, S., Wolstenholme, C., Reidy J., & Reaney-Wood, S. (2018). *ScratchMaths. Evaluation report and executive summary*. London: Education Endowment Foundation.

*Cavalluzzo, L., Geraghty, T. M., Steele, J. L., Holian, L., Jenkins, F., Alexander, J. M., & Yamasaki, K. Y. (2014). *"Using data" to inform decisions: How teachers use data to inform practice and improve student performance in mathematics. Results from a randomized experiment of program efficacy.* Arlington, VA: CNA Corporation.

Cheung, A., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. Educational Research Review, 9, 88-113.

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

Chow, J. C., & Ekholm, E. (2018). Do published studies yield larger effect sizes than unpublished studies in education and special education? A meta-review. *Educational Psychology Review, 30*(3), 727–744. doi:10.1007/s1064801894377

*Clarke, B., Baker, S., Smolkowski, K., Doabler, C., Strand Cary, M., & Fien, H. (2015). Investigating the efficacy of a core kindergarten mathematics curriculum to improve student mathematics learning outcomes. *Journal of Research on Educational Effectiveness, 8*(3), 303–324. doi:10.1080/19345747.2014.980021

*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), 2332858417706899.

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

*Clarke, B., Doabler, C. T., Strand Cary, M., Kosty, D., Baker, S., Fien, H., & Smolkowski, K. (2014). Preliminary evaluation of a tier 2 mathematics intervention for first-grade students: Using a theory of change to guide formative evaluation activities. *School Psychology Review, 43*(2), 160–178.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 294–343.

*Connor, C. M., Mazzocco, M. M., Kurz, T., Crowe, E. C., Tighe, E. L., Wood, T. S., & Morrison, F. J. (2018). Using assessment to individualize early mathematics instruction. *Journal of School Psychology, 66*, 97–113. doi: 10.1016/j.jsp.2017.04.005

de Boer, H., Donker, A. S., & van der Werf, M. P. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research, 84*(4), 509–545. doi:10.3102/0034654314540006

Desimone, L. M., & Garet, M. S. (2015). Best practices in teachers' professional development in the United States. *Psychology, Society and Education, 7*(3), 252–263.

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. doi:10.1177/0014402916660084

*Dominguez, P. S., Nicholls, C., & Storandt, B. (2006). *Experimental methods and results in a study of PBS TeacherLine Math Courses*. Syracuse, NY: Hezel Associates.

*Eddy, R. M., Hankel, N., Hunt, A., Goldman, A., & Murphy, K. (2014). *Houghton Mifflin Harcourt GO Math! efficacy study year two final report.* La Verne, CA: Cobblestone Applied Research & Evaluation, Inc.

*Educational Research Institute of America (2010). *A study of the Singapore math program, Math in Focus, state test results* (Report # 404). Houghton Mifflin Harcourt.

*Educational Research Institute of America (2013). *A study of the instructional effectiveness of Math in Focus* (Report Number 466). Houghton Mifflin Harcourt.

*Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L.,...Bryant, J. D. (2013a). Effects of first-grade number knowledge tutoring with

contrasting forms of practice. *Journal of Educational Psychology*, *105*(1), 58–77. doi:10.1037/a0030127

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

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

*Fuchs, L. S., Schumacher, R. F., Long, J., Jessica, N., Malone, A. S., Amber, W., & ... Changas, P. (2016a). Effects of intervention to improve at-risk fourth graders' understanding, calculations, and word problems with fractions. *Elementary School Journal*, *116*(4), 625–651. doi:10.1080/19345747.2015.1123336

*Fuchs, L. S., Schumacher, R. F., Long, J., Namkung, J., Hamlett, C. L., Cirino, P. T., ... & Changas, P. (2013b). Improving at-risk learners' understanding of fractions. *Journal of Educational Psychology*, *105*(3), 683–700.

Ganley, C. M., & Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. *Learning and Individual Differences*, *47*, 182–193. doi:10.1016/j.lindif.2016.01.002

*Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., & Borman, G. D. (2016). *Focusing on mathematical knowledge: The impact of content-intensive*

teacher professional development (NCEE 2016-4010). Washington, DC: U.S. Department of Education.

*Gatti, G. G. (2009). *Pearson SuccessMaker math pilot study. 2008-09 final report*. Pittsburgh, PA: Gatti Evaluation Inc.

*Gatti, G. (2013). *Pearson SuccessMaker response to intervention study: Final report*. Pittsburgh, PA: Gatti Evaluation. Inc.

*Gatti, G., & Giordano, K. (2008). *Pearson Investigations in Number, Data, & Space efficacy study: 2007-08 School Year Report*. Pittsburgh, PA: Gatti Evaluation, Inc.

*Gatti, G. G., & Petrochenkov, K. (2010). *Pearson SuccessMaker math efficacy study: 2009–10 final report*. Pittsburgh, PA: Gatti Evaluation Inc.

*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. doi:10.3102/0002831214565787

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches (REL 2014-010)*. Retrieved from http://ies.ed.gov/ncee/edlabs.

*Gorard, S., Siddiqui, N., & See, B. H. (2015). *Philosophy for Children. Evaluation report and executive summary*. London: Education Endowment Foundation.

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

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. https://doi.org/10.1002/jrsm.5

*Heller, J. I. (2010). *The impact of Math Pathways & Pitfalls on students' mathematics achievement and mathematical language development: A study conducted in schools with high concentrations of Latino/a students and English learners.* San Francisco, CA: WestEd.

Higgins, K., Huscroft-D'Angelo, J., & Crawford, L. (2019). Effects of technology in mathematics on achievement, motivation, and attitude: A meta-analysis. *Journal of Educational Computing Research, 57*(2), 283–319.

*Hodgen, J., Adkins, M., Ainsworth, S., & Evans, S. (2019). *Catch Up® Numeracy. Evaluation Report and Executive Summary*. London: Education Endowment Foundation.

*Jaciw, A. P., Hegseth, W. M., Lin, L., Toby, M., Newman, D., Ma, B., & Zacamy, J. (2016). Assessing impacts of Math in Focus, a "Singapore Math" program. *Journal of Research on Educational Effectiveness, 9*(4), 473–502. doi:10.1080/19345747.2016.1164777

*Jacob, R., Hill, H., & Corey, D. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research on Educational Effectiveness, 10* (2), 379–407. doi: 10.1080/19345747.2016.1273411

Jacobse, A. E., & Harskamp, E. G. (2011). *A meta-analysis of the effects of instructional interventions on students' mathematics achievement*. Groningen: GION, Gronings Instituut voor Onderzoek van Onderwijs, Opvoeding en Ontwikkeling, Rijksuniversiteit Groningen.

*Jordan, J. (2009). *Math Connects: National field study: Student learning, student attitudes and teachers' reports on program effectiveness: Evaluation report*. Cincinnati, OH: University of Cincinnati Evaluation Services Center.

*Karper, J., & Melnick, S. A. (1993). The effectiveness of Team Accelerated Instruction on high achievers in mathematics. *Journal of Instructional Psychology*, *20*(1), 49–54.

*Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis*, *35*(4), 481–499. doi:10.3102/0162373713498930

*Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness*, *9*(sup1), 188–208. doi:10.1080/19345747.2015.1116031

*Kutaka, T. S., Smith, W. M., Albano, A. D., Edwards, C. P., Ren, L., Beattie, H. L.,...Stroup, W. W. (2017). Connecting teacher professional development and student mathematics achievement: Mediating belonging with multimodal explorations in language, identity, and culture. *Journal of Teacher Education*, *68*(2), 140–154. doi:10.1177/0022487116687551

*Lambert, R., Algozzine, B., & McGee, J. (2014). Effects of progress monitoring on math performance of at-risk students. *British Journal of Education, Society and Behavioural Science*, *4*(4), 527–540.

*Lehmann, R. H., & Seeber, S. (2005). *Accelerated Math in grades 4 through 6: Evaluation of an experimental program in 15 schools in North Rhine-Westphalia*. Berlin: Humboldt University.

*Lenard, M., & Rhea, A. (2019). *Adaptive Math and Student Achievement: Evidence from a Randomized Controlled Trial of DreamBox Learning*. Paper presented ar the Annual Meeting of the Society for Research in Effective Education, Washington, DC.

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review, 22*(3), 215–243. doi:10.1007/s10648-010-9125-8

Lipsey, M. W. (2019). Identifying potentially interesting variables and analysis opportunities. In *The Handbook of Research Synthesis and Meta-Analysis* (3rd ed., pp. 141–151). New York: Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

*Lloyd, C., Edovald, T., Morris, S., Kiss, Z., Skipp, A., & Haywood, S. (2015). *Durham shared maths project. Evaluation report and Executive summary.* London: Education Endowment Foundation.

Lynch, K., Hill, H. C., Gonzales, K.-L., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41* (3), 260-303.

*Magnolia Consulting (2012). *A final report for the evaluation of Pearson's Waterford Early Learning program: Year 2*. Charlottesville, VA: Magnolia Consulting.

*Malone, A. S., Fuchs, L. S., Sterba, S. K., Fuchs, D., & Foreman-Murray, L. (2019). Does an integrated focus on fractions and decimals improve at-risk students' rational number magnitude performance?. *Contemporary Educational Psychology, 59*, 101782. doi: 10.1016/j.cedpsych.2019.101782

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*(7), e1000097. doi:10.1371/journal.pmed1000097

*Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G., & Barton, A. (2016). *ReflectED. Evaluation report and executive summary*. London: Education Endowment Foundation.

Muijs, D., & Bokhove, C. (2020). *Metacognition and self-regulation: Evidence review.* London: Education Endowment Foundation.

National Center for Educational Statistics (2019). *National Assessment of Educational Progress.* Washington, DC: National Center for Educational Statistics.

Neitzel, A., Lake, C., Pellegrini, M., & Slavin, R. (2020). A synthesis of quantitative research on programs for struggling readers in elementary schools. Available at www.bestevidence.org. Manuscript submitted for publication.

*Newman, D., Finney, P. B., Bell, S., Turner, H., Jaciw, A. P., Zacamy, J. L., & Gould, L. F. (2012). *Evaluation of the effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI).* (NCEE 2012–4008). Washington, DC: U.S. Department of Education.

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

*Nunes, T., Malmberg, L., Evans, D., Sanders-Ellis, D., Baker, S., Barros, R., Bryant, P., & Evangelou, M. (2019). *onebillion. Evaluation Report*. London: Education Endowment Foundation.

OECD (2019). *PISA 2018 Results (Volume I): What Students Know and Can Do.* Paris: OECD Publishing https://doi.org/10.1787/5f07c754-en.

*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. doi: 10.1080/19345747.2019.1571653

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

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86*(1), 207–236. https://doi.org/10.3102/0034654315582067

*Prast, E. J., Van de Weijer-Bergsma, E., Kroesbergen, E. H., & Van Luit, J. E. (2018). Differentiated instruction in primary mathematics: Effects of teacher professional development on student achievement. *Learning and Instruction, 54*, 22–34.

Pustejovsky, J. (2020). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections* (Version R package version 0.4.1) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

*Randel, B., Apthorp, H., Beesley, D., Clark, F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *The Journal of Educational Research, 109*(5), 491–502. doi:10.1080/00220671.2014.992581

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

*Reid, E. E., Chen, J. Q., & McCray, J. (2014). *Achieving high standards for Pre-K-Grade 3 mathematics: A whole teacher approach to professional development*. Paper presented at the Annual Meeting of the Society for Research on Effective Education, Washington, DC.

*Resendez, M., & Azin, M. (2006). *2005 Scott Foresman–Addison Wesley Elementary Math randomized control trial: Final report.* Jackson, WY: PRES Associates, Inc.

*Resendez, M., Azin, M., & Strobel, A. (2009). *A study on the effects of Pearson's 2009 enVision math program: Final summative report*. Jackson, WY: Press Associates.

*Resendez, M., & Manley, M. A. (2005). *Final report: A study on the effectiveness of the 2004 Scott Foresman–Addison Wesley Elementary Math program*. Jackson, WY: PRES Associates.

*Rosen, Y., & Beck-Hill, D. (2012). Intertwining digital content and a one-to-one laptop environment in teaching and learning: Lessons from the Time to Know program. *Journal of Research on Technology in Education, 44*(3), 225–241. doi:10.1080/15391523.2012.10782588

*Roy, P., Rutt, S., Easton, C., Sims, D., Bradshaw S., & McNamara, S. (2019). *Stop and Think: Learning Counterintuitive Concepts. Evaluation Report and Executive Summary*. London: Education Endowment Foundation

*Rudd, P., Berenice, Villaneuva Aguilera A. B., Elliott, L., Chambers, B. (2017). *MathsFlip: Flipped Learning. Evaluation Report and Executive Summary*. London: Education Endowment Foundation.

*Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J.,…Martinez, E. (2014). A randomized trial of an elementary school mathematics software intervention:

Spatial-Temporal Math. *Journal of Research on Educational Effectiveness, 7*(4), 358–383. doi:10.1080/19345747.2013.856978

*Rutt, S., Easton, C., & Stacey, O. (2014). *Catch Up® Numeracy: Evaluation report and executive summary*. London, UK: Education Endowment Foundation.

Savelsbergh, E. R., Prins, G. T., Rietbergen, C., Fechner, S., Vaessen, B. E., Draijer, J. M., & Bakker, A. (2016). Effects of innovative science and mathematics teaching on student attitudes and achievement: A meta-analytic study. *Educational Research Review, 19*, 158–172. doi:10.1016/j.edurev.2016.07.003

*Schoen, R. C., LaVenia, M., & Tazar, A. M. (2018, March). *Effects of a two-year Cognitively Guided Instruction professional development program on first- and second-grade student achievement in mathematics.* Paper presented at the annual meeting of the Society for Research in Effective Education, Washington, DC.

*Schwarz, P. (2019). Raising the Bar District-Wide Using Symphony Math. Symphony Learning: Research Study. Retrieved from https://content.symphonylearning.com/assets/web/SLC_Graves_2020_02_27.pdf

*See, B. H., Morris, R., Gorard S. G., Siddiqui, N. (2018). *Maths Counts. Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Maths_Counts.pdf

*Shechtman, N., Roschelle, J., Feng, M., & Singleton, C. (2019). An Efficacy Study of a Digital Core Curriculum for Grade 5 Mathematics. *AERA Open, 5*(2), doi: 0.1177/2332858419850482

Slavin, R. E. (2017). Instruction based on cooperative learning. In R. Mayer & P. Alexander (Eds.), Handbook of research on learning and instruction. New York: Routledge.

Slavin, R., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. Review of Educational Research, 78 (3), 427-515.

*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. doi:10.3102/0002831212469045

*Solomon, T., Martinussen, R., Dupuis, A., Gervan, S., Chaban, P., Tannock, R., & Ferguson, B. (2011). *Investigation of a cognitive science based approach to mathematics instruction.* Paper presented at the Biennial Meeting of the Society for Research in Child Development, Montreal, Canada.

*Stevens, R. J., & Slavin, R. E. (1995). The cooperative elementary school: Effects on students' achievement, attitudes, and social relations. *American Educational Research Journal, 32*(2), 321–351. doi:10.3102/00028312032002321

*Stokes, L., Hudson-Sharp, N., Dorsett, R., Rolfe, H., Anders, J., George, A., Buzzeo, J., & Munro-Lott, N. (2018). *Mathematical Reasoning. Evaluation report and executive summary*. London: Education Endowment Foundation.

*Strobel, A., Resendez, M., & DuBose, D. (2017). *enVisionmath2.0 Year 2 RCT Study Final Report.* Thayne, WY: Strobel Consulting, LLC.

*Styers, M. & Baird-Wilkerson, S. (2011). *A final report for the evaluation of Pearson's focusMATH Program*. Charlottesville, VA: Magnolia Consulting.

*Suppes, P., Holland, P. W., Hu, Y., & Vu, M.T. (2013). Effectiveness of an individualized computer-driven online math K-5 course in eight California Title I elementary schools. *Educational Assessment, 18*(3), 162–181. doi:10.1080/10627197.2013.814516

*Sutherland, A., Broeks, M., Sim, M., Brown, E., Iakovidou, E., Ilie, S., Jarke, H., Belanger, J. (2019). *Digital Feedback in Primary Maths*. *Evaluation Report and Executive Summary*. London: Education Endowment Foundation.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology, 2,* 85-112.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

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

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

*Torgerson, C. J., Bell, K., Coleman, E., Elliott, L., Fairhurst, C., Gascoine, L., Hewitt, C.E., & Torgerson, D. J. (2018). *Tutor Trust: Affordable Primary Tuition*. *Evaluation report and executive summary*. London: Education Endowment Foundation.

*Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., de Castilla, V. R., & Sullivan, K. (2015). *Preliminary findings from a multi-year scale-up effectiveness trial of Everyday Mathematics*. Paper presented at the Society for Research on Effective Education, Washington, DC.

Valentine, J. C., Hedges, L. V., & Cooper, H. M. (2019). *The handbook of research synthesis and meta-analysis (3rd ed.).* Russell Sage Foundation.

*VanDerHeyden, A. M., McLaughlin, T., Algina, J., & Snyder, P. (2012). Randomized evaluation of a supplemental grade-wide mathematics intervention. *American Education Research Journal, 49*, 1251–1284. doi:10.3102/0002831212462736

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. doi:10.18637/jss.v036.i03

*Vignoles, A., Jerrim, J., & Cowan, R. (2015). *Mathematics Mastery: Primary evaluation report*. London: Education Endowment Foundation.

*Wang, H., & Woodworth, K. (2011a). *Evaluation of Rocketship Education's use of DreamBox Learning's online mathematics program*. Menlo Park, CA: SRI International. Retrieved

*Wang, H., & Woodworth, K. (2011b). *A randomized controlled trial of two online mathematics curricula*. Paper presented at the Annual Meeting of the Society for Research on Effective Education, Washington, DC.

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of tier 2 type reading interventions in grades K-3. *Educational Psychology Review, 28*(3), 551–576. doi:10.1007/s10648-015-9321-7

*Weis, R., Osborne, K. J., & Dean, E. L. (2015). Effectiveness of a universal, interdependent group contingency program on children's academic achievement: A

countywide evaluation. *Journal of Applied School Psychology*, *31*(3), 199–218. doi:10.1080/15377903.2015.1025322

*West, M. R., Morton, B. A., & Herlihy, C. M. (2016). *Achievement Network's Investing in Innovation expansion: Impacts on educator practice and student achievement*. Cambridge, MA: Center for Educational Policy Research, Harvard University.

What Works Clearinghouse. (2020). *Standards handbook* (Version 4.1). Washington, DC: What Works Clearinghouse.

*Wijekumar, K., Hitchcock, J., Turner, H., Lei, P. W., & Peck, K. (2009). *A multisite cluster randomized trial of the effects of CompassLearning Odyssey® Math on the math achievement of selected Grade 4 students in the Mid-Atlantic region* (NCEE 2009-4068). Washington, DC: U.S. Department of Education.

*Worth, J., Sizmur, J., Ager, R., & Styles, B. (2015). *Improving numeracy and literacy*. London: Education Endowment Foundation.

*Wright, W., Dorsett, R., Anders, J., Buzzeo, J., Runge, J., & Sanders, M. (2019). *Improving Working Memory*. *Evaluation Report and Executive Summary*. London: Education Endowment Foundation.

*Ysseldyke, J., & Bolt, D. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*(3), 453–467.

Table 1. Meta-regression results.

Coefficient	Reference group	beta	SE	t	df	p
Null Model						
Intercept		0.11	0.02	6.22	71.62	0.000
Meta-Regression						
Intercept	Tutoring	0.09	0.01	7.62	36.84	0.000
PD Focused on Mathematics Content and Pedagogy		-0.12	0.07	-1.70	22.93	0.103
PD Focused on Classroom Management, Motivation, and Cognition		0.05	0.08	0.61	15.65	0.554
PD Focused on Implementation of Traditional and Digital Curricula		-0.14	0.06	-2.16	9.99	0.056
Traditional and Digital Curricula with Limited Professional Development		-0.12	0.07	-1.81	18.41	0.086
Benchmark Assessments		-0.15	0.10	-1.51	7.04	0.173
PD Focused on Implementation of Traditional Curricula	PD Focused on Implementation of Digital Curricula	0.11	0.04	2.56	7.27	0.036
Digital Curricula	Traditional Curricula	0.04	0.04	0.93	23.26	0.363
Quasi-Experiments	Randomized Studies	0.10	0.04	2.39	12.48	0.033
K-2	Mixed	-0.01	0.03	-0.35	19.00	0.730
3-6		0.00	0.03	0.01	16.21	0.996
Low Achievers	Mixed Achievers	0.06	0.03	2.02	10.60	0.069
Moderate/high Achievers		-0.01	0.02	-0.60	11.60	0.557
Low SES	Mixed SES	-0.01	0.03	-0.42	31.22	0.674
Moderate/high SES		0.02	0.03	0.83	31.68	0.413
International Studies	U.S. Studies	-0.03	0.03	-0.81	26.81	0.426
One-to-Small Group Tutoring	One-to-One Tutoring	0.11	0.08	1.41	15.09	0.179

Note. Meta-regression model also controlled for cross-age and online tutoring.

Table 2. Mean Effect Sizes of Program Categories and Subcategories.

Table	Category	k	n	ES	SE	t	df	p
3	Tutoring programs	21	35	+0.20	0.05	4.27	7.90	0.003
	One-to-One Tutoring	8	13	+0.19	0.05	3.53	7.58	0.008
	One-to-Small Group Tutoring	13	22	+0.30	0.05	5.47	12.44	0.000
4	Professional Development Focused on Mathematics Content and Pedagogy	9	18	+0.03	0.03	0.96	8.09	0.367
5	Professional Development Focused on Classroom Management, Motivation, and Cognition	8	18	+0.19	0.04	4.36	5.74	0.005
6	Professional Development Focused on Implementation of Traditional and Digital Curricula	12	33	+0.02	0.03	0.60	3.10	0.590
	Professional Development Focused on Implementation of Traditional Curricula	7	16	+0.12	0.02	5.09	5.69	0.003
	Professional Development Focused on Implementation of Digital Curricula	5	17	+0.01	0.03	0.23	3.10	0.832
7	Traditional and Digital Curricula With Limited Professional Development	29	62	+0.03	0.03	1.12	13.99	0.282
	Traditional Curricula	15	31	+0.03	0.04	0.73	13.54	0.475
	Digital Curricula	14	31	+0.07	0.02	3.12	10.66	0.010
8	Benchmark Assessments	4	5	0.00	0.08	-0.01	3.20	0.994

Note. k=number of studies; *n* = number of outcomes; ES=effect size; SE=standard error; df=degrees of freedom

Table 3. *Tutoring Programs*

Study	Design	Duration	Sample Size	Grade	Sample Characteristics	Posttest	Effect size	Study ES
							Category Mean: +0.20*	
				One-to-one Tutoring			**Subcategory Mean: +0.19***	
				One-to-one Tutoring by Teachers				
Math Recovery								
Smith et al. (2013)	QE	1 year	775 students (259E, 516C)	1	48% minority, 15% ELL, 65% FRL.	WJ-Math Fluency	+0.15*	+0.24*
						WJ-App. Problems	+0.28*	
						WJ-Quant Concepts	+0.24*	
						WJ-Math Reasoning	+0.30*	
Numbers Count								
Torgerson et al. (2013)	SR	12 weeks	418 students (144E, 274C)	Year 2 (Grade 1)	England. 75% FRL.	Progress in Math (PIM 6)		+0.33*
				One-to-one Tutoring by Teaching Assistants				
Catch Up® Numeracy							**Program Mean: +0.05**	
Hodgen et al. (2019)	CR	1 year	142 schools 1481 students (737E, 744C)	Year 4, 5 (Grade 3, 4)	Urban and rural schools in England. 22% FRL.	Progress Test in Mathematics		-0.04
Rutt et al. (2014)	SR	30 weeks	216 students (108E, 108C)	Year 2-6 (Grade 1-5)	England. 35% FRL.	Progress Test in Mathematics		+0.21*
Galaxy Math								
Fuchs et al. (2013a)	SR	16 weeks	591 students (385E, 206C)	1	Southeast school district. 69% AA, 7% H, 83% FRL.	Word Problems		+0.25*
Maths Counts								
See et al. (2018)	SR	3 months	291 students (147E, 144C)	Year 3-6 (grade 2-5)	Low performing students in England. 37% FRL, 54% SEN.	Key Stage 2		+0.11

Pirate Math							
Fuchs et al. (2010)	SR	16 weeks	150 students (100E, 50C)	3	Nashville and Houston. 35% SPED, 19% ELL, 75% FRL, 56% AA, 29% H.		+0.37*
One-to-one Tutoring by Paid Volunteers							
MathCorps							
Parker et al. (2019)	SR	6 months	284 students (183E, 101C)	4-6	Minnesota. 35%W, 27%AA, 20% A, 61%FRL.	STAR Math	+0.20*
One-to-Small Group Tutoring							**Subcategory Mean: +0.30***
One-to-Small Group Tutoring by Teachers							
Number Rockets							
Gersten et al. (2015)	CR	6 months	76 schools 994 students (615E, 379C)	1	44% AA, 46% H, 34% FRL.	TEMA-3	+0.34*
One-to- Small Group Tutoring by Teaching Assistants							
1stClass@Number							
Nunes et al. (2018)	CR	3 months	122 schools 503 students (251E, 252C)	Year 2 (grade 1)	Schools in England. 40% FRL	Key Stage 1	+0.01
Affordable Primary Tuition							
Torgerson et al. (2018)	CR	12 weeks	102 schools 1201 students (567E, 634C)	Year 6 (Grade 5)	England. 48% FRL, 72%W.	Key Stage 2	+0.19
FocusMATH							
Styers & Baird-Wilkerson (2011)	SR	1 year	341 students (166E, 175C)	3, 5	23% AA, 33% H, 24% ELL, 12% SPED, 71% FRL	KeyMath 3	+0.24*
Fraction Face-Off!							**Program Mean: +0.57***
Fuchs et al. (2013b)	SR	12 weeks	259 students (129E, 130C)	4	82% FRL, 11% ELL, 53% AA, 25% W, 19% H.	NAEP Items	+0.88*
Fuchs et al. (2016a)	SR	12 weeks	213 students (143E, 70C)	4	17% ELL, 88% FRL, 15% SPED, 58% AA, 16% W, 17% H	NAEP Items	+0.39*

Fuchs et al. (2016b)	SR	12 weeks	212 students (142E, 70C)	4	49% AA, 27% H, 18% ELL, 90% FRL.	NAEP Items		+0.64*
Malone et al. (2019)	SR	12 weeks	225 students (149E, 76C)	4	16% W, 43% AA, 25% H, 20% ELL, 88% FRL.	NAEP Items		+0.29*
Fusion Math								
Clarke et al. (2014)	SR	19 weeks	78 students (38E, 40C)	1	Pacific Northwest. 20% H, 18% ELL, 70% FRL, 12% SPED.	SAT-10		+0.11
Onebillion Maths Apps								
Nunes et al. (2019)	CR	12 weeks	112 schools 1089 students (543E, 546C)	Year 1 (K)	England. 25% FRL	PTM		+0.24*
ROOTS							**Program Mean: +0.19***	
Clarke et al. (2016)	SR	4 months	290 students (203E, 87C)	K	Oregon. 5% AA, 58% W, 33% H, 32% LEP, 11% SPED	TEMA-3	+0.32*	+0.16
						NSB	+0.16	
						SESAT	+0.001	
Doabler et al. (2016)	SR	5 months	292 students (208E, 82C)	K	Boston. 7% AA, 89% W, 50% H, 26% ELL.	TEMA-3	+0.31*	+0.32*
						NSB	+0.40*	
						SESAT	+0.24	
Clarke et al. (2017)	SR	4 months	689 students (527E, 162C)	K	Oregon. 55% W, 26% H, 26% LEP, 87% FRL.	TEMA-3	+0.25*	+0.15
						NSB	+0.09	
						SESAT	+0.12	
Online One-to-one Tutoring								
Affordable Online Maths Tuition								
Torgerson et al. (2016)	CR	27 weeks	64 schools 578 students (289E, 289C)	Year 6 (Grade 5)	England. 92% FRL, 43% minority.	Key Stage 2		-0.03
Cross-age Peer Tutoring								
Shared Maths								
Lloyd et al. (2015)	CR	2 years	79 schools Year 3 (tutees) 2786 students Year 5 (tutors) 2683 students	Year 3, 5 (Grades 2, 4)	England. 22% FRL, 86% W, 4% AA, 5% A.	ICAS-Year 3	+0.01	+0.02
						ICAS-Year 5	+0.02	

Note for Tables 3-8.
Design/Treatment: SR=Student Randomized, CR=Cluster Randomized, QE=Quasi Experiment, CQE=Cluster Quasi-Experiment
Measures: BAM: Balanced Assessment in Mathematics, CAT: California Achievement Test, CMT-Math: Connecticut Mastery Test, CST: California Standards Test, CSAP: Colorado Student Assessment Program, ECLS-K: Early Childhood Longitudinal Program, FCAT: Florida Comprehensive Assessment Test, GMADE: Group Mathematics Assessment and Diagnostic Evaluation, HCPS II: Hawaii Content and Performance Standards, ICAS: Interactive Computerised Assessment Systemin, CAS: Interactive Computerized Assessment System, ISAT: Illinois Student Achievement Test, ISTEP+: Indiana State Test of Educational Proficiency, ITBS: Iowa Test of Basic Skills, MAP: Measure of Academic Progress, MAT- Metropolitan Achievement Test, MEAP: Michigan Educational Assessment Program, NAEP: National Assessment of Educational Progress, NJASK: New Jersey State Test; NSB: Brief Number Sense Screener, Nevada CRT: Nevada Criterion Referenced Test, NWEA: Northwest Evaluation Association, PTM: Progress Test in Maths, SAT 10: Stanford Achievement Test 10, SESAT: Stanford Early School Achievement Test; SOL: Virginia Standards of Learning, STAR Math: Standardized Testing and Reporting, TAKS: Texas Assessment of Knowledge and Skills, TEMA-3: Test of Early Mathematics Ability 3, WJ III: Woodcock-Johnson III.
Demographics: A=Asian, AA=African-American, H=Hispanic, W=White, FRL=Free/Reduced Lunch, ELL=English Language Learner, LD=Learning Disabilities, SPED=Special Education.
* $p < .05$ at the appropriate level of analysis (cluster or individual).

Table 4. *Professional Development Focused on Mathematics Content and Pedagogy*

Study	Design	Duration	Sample Size	Grade	Sample Characteristics	Posttest	Effect size	Study ES
							Category Mean: +0.03	
CASL								
Randel et al. (2016)	CR	1-2 years	67 schools 9,596 students (4,420E, 5,176C)	4,5	CO. 56% W, 27% H, 47% FRL.	CSAP		+0.01
Cognitively Guided Instruction								
Schoen et al. (2018)	CR	2 years	22 schools 2,230 students (1,110 E, 1,120C)	1, 2	37%W, 37% H, 18% AA, 22% ELL, 60% FRL	ITBS		0.00
						Grade 1 Comp.	-0.08	
						Grade 1 Problems	+0.09	
						Grade 2 Comp.	-0.07	
						Grade 2 Problems	+0.06	
Intel Math								
Garet et al. (2016)	CR	1 year	165 teachers 3,677 students (1,760E, 1,917C)	4	46% W, 14% AA, 30% H, 58% FRL, 12% ELL, 14% SPED.	State tests	-0.06*	-0.05*
						NWEA	-0.05	
Math Solutions								
Jacob et al. (2017)	CR	2 years	74 classes 1,453 students (727E, 726C)	4, 5	63% AA, 21% W, 14% SPED	State tests		
						Grade 4	+0.04	+0.06
						Grade 5	+0.08	
PBS TeacherLine								
Dominguez et al. (2006)	CR	1 year	87 teachers 1,119 students (523E, 596C)	3-5	FL, SC, NY.	Algebra test	-0.02	+0.03
						Geometry test	+0.08	
Philosophy for Children								
Gorard et al. (2015)	CR	1 year	48 schools 1,529 students (772E, 757C)	Year 5 (Grade 4)	England. 47% FRL, 19% SPED, 12% ELL, 26% minority.	Key Stage 2		+0.10
Primarily Math								
Kutaka et al. (2017)	CQE	1 year	218 teachers 809 students (313E, 496C)	K-2	3 urban school districts.	TEMA-3		+0.14
Project GROW								

Prast et al. (2018)	CR	1 year	30 schools 3,514 students	1-6	Schools from Netherlands.	Cito Mathematics Test	+0.11*
Using Data							
Cavalluzzo et al. (2014)	CR	2 years	59 schools 10,877 students (5,384E, 4,903C)	4,5	FL. 47% AA, 9% H, 66% FRL, 10% SPED.	FCAT	+0.01

Table 5. *Professional Development Focused on Classroom Management, Motivation, and Cognition*

Study	Design	Duration	Sample Size	Grade	Sample Characteristics	Posttest	Effect size	Study ES
							Category Mean: +0.19*	
Individualized Student Instruction (ISI)								
Connor et al. (2018)	CR	1 year	32 teachers 370 students (205E, 165C)	2	North FL. 84%W, 5% AA	Woodcock Math Fluency	+0.16	+0.11
						Key Math	+0.07	
PAX Good Behavior Game								
Weis et al. (2015)	CQE	1 year	49 classes 703 students (402E, 301C)	1, 2	Ohio. 82% W, 48% FRL	MAP		+0.32*
ReflectEd								
Motteram et al. (2016)	CR	1 year	65 classes 1570 students (839E, 731C)	Year 5 (Grade 4)	England	InCAS		+0.32
Spring Math								
VanDerHayden et al. (2012)	CR	1 year	23 classes 187 students (106E, 81C)	5	Mississippi. 34%W, 36%AA, 11% SPED, 57% FRL	State Test		-0.05
Stop and Think								
Roy et al. (2019)	CR	1 year	84 year groups 2702 students (1343E, 1359C)	Year 3, 5 (grade 2, 4)	England. 30% FRL	Progress Test in Maths (PTM)		+0.09
TAI							**Program Mean: +0.11**	
Stevens & Slavin (1995)	CQE	2 years	5 schools 873 students (411E, 462C)	2-6	MD. 7% minority, 10% FRL, 9% SPED	CAT-Computation	+0.29	+0.24
						CAT-Application	+0.20	
Karper & Melnick (1993)	CQE	1 year	8 classes 165 students (84E, 81C)	4-5	Hershey, PA.	District Test		
						Grade 4	-0.05	-0.09
						Grade 5	-0.12	
Working Memory								
Wright et al. (2019)	CR	5 months	171 schools 1822 students (882E, 940C)	Year 3 (grade 2)	England. 37% FRL, 80% W	GL Assessment British Ability		+0.22

Table 6. *Professional Development Focused on Implementation of Traditional and Digital Curricula*

Study	Design	Duration	Sample Size	Grade	Sample Characteristics	Posttest	Effect size	Study ES
							Category Mean: +0.02	
Professional Development Focused on Implementation of Traditional Curricula							**Subcategory Mean: +0.12***	
AMSTI								
Newman et al. (2012)	CR	1 year	40 schools 9,370 students (5,111E, 4,259C)	4-5	49% minority, 64% FRL.	SAT 10		+0.05
EarlyMath								
Reid et al. (2014)	CQE	2 years	16 schools 903 students (443, 460C)	K-2	Midwestern city.	WJ-Applied Problems		+0.01
Math Pathways & Pitfalls								
Heller (2010)	CR	1 year	121 classes 2,160 students (1,204E, 956C)	4, 5	AZ, CA, IL. 55% ELL, 76% FRL, 8% AA, 69% H, 9% W.	State tests Grade 4 Grade 5	 +0.04 +0.08	+0.06
Mathematics Mastery								
Vignoles et al. (2015)	CR	1 year	83 schools 4,176 students (2,160E, 2,016C)	Year 1 (Grade K)	Schools across England.	Number Knowledge Test		+0.10
Mathematics Reasoning							**Program Mean: +0.10**	
Stokes et al. (2018)	CR	12 weeks	160 schools 6,353 students (3,238E, 3,115C)	Year 2 (Grade 1)	England. 23% FRL	Progress in Math (PIM 7)		+0.08
Worth et al. (2015)	CR	4 months	36 schools 1,365 students (517E, 848C)	Year 2 (Grade 1)	England. 16% FRL, 14% SPED, 14% ELL.	Progress in Math (PIM 7)		+0.20*
Math Expressions								
Agodini et al. (2010)	CR	1 year	90 schools 4,114 students (2,036E, 2,078C)	1, 2	CT, FL, KY, MN, MS, MO, NY, NV, SC, TX. 26% AA, 30% H, 10% ELL.	ECLS-K Grade 1 Grade 2	 +0.11* +0.12*	+0.11*

Professional Development Focused on Implementation of Digital Curricula							**Subcategory Mean: +0.01**
MathsFlip							
Rudd et al. (2017)	CR	1 year	24 schools 1,129 students (542E, 587)	Year 5, 6 (grade 4, 5)	England. 25% FRL, 37% ELL	Key Stage 2	+0.07
Odyssey Math							
Wijekumar et al. (2009)	CR	1 year	122 teachers 2,456 students (1,223E, ,233C)	4	DE, NJ, PA. 18% FRL, 25% minority, 7% ELL.	TerraNova	+0.02
Reasoning Mind							**Program Mean: -0.04**
Shechtman et al. (2019)	CR	1 year	46 schools 1,921 students (941E, 980C)	5	Urban, rural and suburban schools in West Virginia. 94% W, 50% FRL	WVGSA	-0.13
Wang & Woodworth (2011b)	SR	4 months	651 students (521E, 130C)	2-5	San Francisco Bay Area. 87% H, 81% ELL, 88% FRL.	NWEA-Math Over. -0.02 NWEA-Probl. Solv. -0.05 NWEA-Num. sense +0.01 NWEA-Comp. -0.08 NWEA-Geometry +0.11 NWEA-Statistics -0.02	-0.01
Time to Know							
Rosen & Beck-Hill (2012)	CQE	6 months	4 schools 476 students (283E, 193C)	4-5	Dallas, TX 18% AA, 63% H	TAKS	+0.31

Table 7. *Traditional and Digital Curricula with Limited Professional Development*

Study	Design	Duration	Sample Size	Grade	Sample Characteristics	Posttest	Effect size	Study ES
							Category Mean:	**+0.03**
Traditional Curricula								
							Subcategory Mean:	**+0.03**
Early Learning in Mathematics								
Clarke et al. (2015)	CR	1 year	129 classes 2,116 students (1,134E, 982C)	K	OR, TX. 56% FRL, 38% ELL, 36% H, 8% SPED.	TEMA-3		+0.11
enVisionMATH / Scott Foresman-Addison Wesley Elementary Math							**Program Mean:**	**-0.02**
Resendez & Azin (2006)	CR	1 year	39 classes 863 students (445E, 418C)	3, 5	OH, NJ 9% AA, 18% FRL.	TerraNova-Math Tot.	-0.07	-0.01
Resendez & Manley (2005)	CR	1 year	35 teachers 645 students (352E, 293C)	2, 4	WA, WY, VA, KY 20% AA, 9% H, 10% ELL, 46% FRL.	TerraNova-Math Tot.	+0.10	-0.05
						TerraNova-Comp.	-0.21	
						TerraNova-Comp.	+0.05	
Resendez et al. (2009)	CR	2 years	44 teachers 659 students (349, 310C)	2-3, 4-5	MT, OH, NH, MA, KY, TN. 95%W, 19% FRL.	MAT-Conc. & Prob. Sol.	-0.13	-0.04
						MAT-Math Comp.	+0.06	
						GMADE	-0.06	
Strobel et al. (2017)	CR	2 years	33 teachers 495 students (285E, 210C)	1-2, 4-5	24% W, 37% AA, 33% H, 15% ELL, 74% FRL.	TerraNova		+0.02
Everyday Mathematics								
Vaden-Kiernan et al. (2015)	CR	2 years	48 schools 4,467 students	K-5	51% AA, 73% FRL.	GMADE		-0.01
GO Math!								
Eddy et al. (2014)	CR	1 year	79 teachers 1,363 students (754E, 609C)	1-3	AZ, ID, IL, MI, OH, PA, UT, 36% AA, 35% H, 31% ELL, 35% FRL.	ITBS		+0.01
Investigations in Number, Data, and Space							**Program Mean:**	**-0.09**
Agodini et al. (2010)	CR	1 year	93 schools 4,019 students (1,941E, 2,078C)	1, 2	CT, FL, KY, MN, MS, MO, NY, NV, SC, TX. 23% AA, 32% H, 13% ELL.	ECLS-K		+0.04
						Grade 1	0.00	
						Grade 2	+0.09	
	CR	1 year	77 classes	1, 4	AZ, MA, OR, SC	GMADE		

Gatti & Giordano (2008)			1,363 students (729E, 634C)		52% FRL, 27% H, 9% AA.	Grade 1 Grade 4	-0.14 -0.31	-0.22*
JUMP Math								
Solomon et al. (2011)	CR	5 months	18 schools 267 students (163E, 104C)	5	Rural Canadian schools, Ontario.	WJ-III		+0.23
Math Connects								
Jordan (2009)	CQE	1 year	139 teachers 1,897 students (844E, 1,053C)	2, 4	61% W, 14% AA, 16% H.	TerraNova Grade 2 Grade 4	 +0.08 -0.04	 +0.02
Math in Focus							**Program Mean: +0.24***	
ERIA (2010)	QE	1 year	678 students (125E, 553C)	4	NJ. 15% FRL, 30% minority, 12% SPED.	NJ ASK		+0.25*
ERIA (2013)	CQE	1 year	33 classes 679 students (362E, 317C)	3	59% minority, 58% FRL, 9% ELL.	ITBS		+0.29
Jaciw et al. (2016)	CR	1 year	18 teams 1,641 students (857E, 784C)	3-5	Clark County, NV. 47% H, 10% AA, 56% FRL, 11% SPED.	SAT10-Probl. Solv SAT10-Procedures Nevada CRT	+0.12* +0.14* +0.05	+0.10
Saxon Math								
Agodini et al. (2010)	CR	1 year	91 schools 4,083 students (2,005E, 2,078C)	1, 2	CT, FL, KY, MN, MS, MO, NY, NV, SC, TX. 21% AA, 40% H, 12% ELL.	ECLS-K Grade 1 Grade 2	 +0.07 +0.17*	 +0.11
Digital Curricula							**Subcategory Mean: +0.07***	
Accelerated Math							**Program Mean: +0.02**	
Lambert et al. (2014)	CR	1 year	36 classes 504 students (256E, 248C)	2-5	Midwestern US. 40% minority, 76% FRL, 18% SPED	TerraNova		+0.02
Lehmann & Seeber (2005)	CQE	4 months	47 classes 1,243 students (577E, 666C)	4-6	Germany. 18% immigrants	Hamburger Schulleistungs-test Grade 4 Grade 5 Grade 6	 +0.01 +0.17 -0.01	 +0.06

Ysseldyke & Bolt (2007)	CR	1 year	36 classes 723 students (368E, 355C)	2-5	AL, FL, SC, TX, MS, MI, NC. 44% AA, 45% H	TerraNova		0.00
Digital Feedback in Primary Maths								
Sutherland et al. (2019)	CR	1 year	108 classes 2133 students (1103E, 1030C)	Year 4, 5 (grade 3, 4)	England. 30% FRL	ACERs Essential Learning Metric (ELM)		-0.04
DreamBox Learning								**Program Mean: +0.10**
Lenard & Rhea (2019)	CR	6 months	24 schools 12,467 students (6,084E, 6,048C)	K-5	School in North Carolina. 18% H, 22% AA, 47% W, 11% LEP, 25% FRL.	Number Knowledge Test (K-2)	+0.12*	+0.08
						North Carolina End-of-Grade EOG (3-5)	+0.03	
Wang & Woodworth (2011a)	SR	4 months	557 students (446E, 111C)	K, 1	San Francisco Area. 87% H, 81% ELL, 88% FRL.	NWEA-Math Over.	+0.11	+0.11
						NWEA-Probl. Solv.	+0.06	
						NWEA-Num. sense	+0.08	
						NWEA-Comp.	+0.13	
						NWEA-Geometry	+0.16*	
						NWEA-Statistics	+0.12	
Educational Program for Gifted Youth (EGPY)								
Suppes et al. (2013)	SR	1 year	1484 students (742E, 742C)	2-5	California. 55% AA, 31% H.	CST		-0.01
ScratchMaths								
Boylan et al. (2018)	CR	2 years	110 schools 5,818 students (2,803E, 3,015C)	Years 5, 6 (Grades 4, 5)	England. 28% FRL.	Key Stage 2		0.00
ST Math								
Rutherford et al. (2014)	CR	1, 2 years	1 yr: 34 schools 10,455 students 2 yrs: 18 schools 2,677 students	3-5	Southern CA. 90% FRL, 85% H, 63% ELL.	CST		+0.08
						1 year	+0.09	
						2 years	+0.03	
SuccessMaker								**Program Mean: +0.08**
Gatti (2009)	CQE	1 year	8 schools 792 students (455E, 337C)	3,5	AZ, FL, MA, NJ. 34% H, 34% FRL, 89% ELL, 47% low achievers.	GMADE		+0.07
						Grade 3	+0.11	
						Grade 5	+0.03	

Gatti (2013)	SR	1 year	490 students (239E, 251C)	5	AZ, CA, KS, MI, OR, TX. 49% H, 8% AA, 11% SPED, 17% LEP, 70% FRL.	GMADE		+0.09
Gatti & Petrochenkov (2010)	CR	1 year	47 classes 913 students (506E, 407C)	3, 5	AZ, AR, CA, IN, KS, PA. 88% ELL, 66% FRL,42% H, 12% AA, 40% low achievers.	GMADE-Grade 3	+0.27	+0.06
						GMADE-Grade 5	-0.19	
Symphony Math								
Schwarz (2019)	CQE	1 year	58 classes 1,202 students (579E, 623C)	1-4	Kentucky. 87% W, 57% FRL.	STAR 360® Math		+0.30
Waterford Early Learning								
Magnolia Consulting (2012)	CR	2 years	57 classes 680 students (425E, 255C)	K-1 1-2	19% AA, 53% H, 17% W, 73% FRL, 32% LEP, 5% SPED.	SAT 10		+0.04

Table 8. *Benchmark Assessments*

Study	Design	Duration	Sample Size	Grade	Sample Characteristics	Posttest	Effect size	Study ES
							Category Mean: 0.00	
Achievement Network (ANet)								
West et al. (2016)	CR	2 years	89 schools 13,233 students (6,617E, 6,616C)	3-5	MA, LA, IL. 87% AA, 15% ELL, 87% FRL.	State tests		-0.09*
Acuity								**Program mean: +0.16**
Konstantopoulos et al. (2013)	CR	1 year	49 schools 11,632 students (5,816E, 5,816C)	3-6	Rural, urban, and suburban schools in IN	ISTEP+		+0.19*
Konstantopoulos et al. (2016)	CR	1 year	55 schools 13,944students (6,972E, 6,972C)	3-6	IN. 53% W, 27% AA, 12% H, 57% FRL 19% SPED.	ISTEP+		+0.13
mClass								
Konstantopoulos et al. (2016)	CR	1 year	55 schools 6,249 students	K-2	IN. 27%AA, 12% H, 57% FRL, 19% SPED.	TerraNova		-0.22*

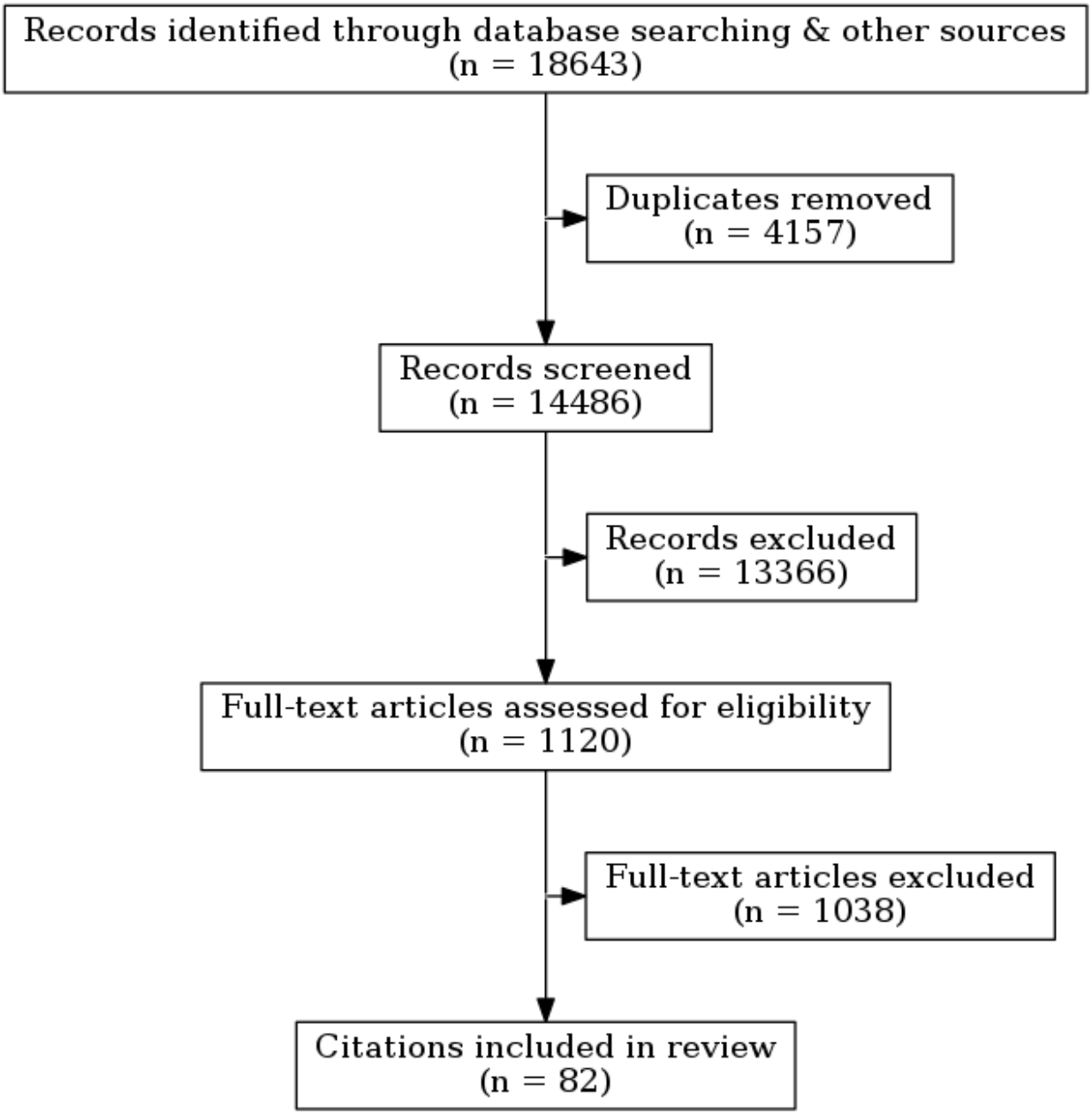
Table 9. Methodological and Substantive Moderators.

Moderator	Level	k	n	ES	SE	t	df	p
Research Design	Quasi-Experiments	13	22	+0.18	0.04	4.67	10.01	0.001
	Randomized studies	72	155	+0.08	0.01	6.21	34.58	0.000
Grade Level	K-2	33	68	+0.09	0.02	4.69	27.22	0.000
	3-6	48	82	+0.10	0.02	4.42	33.01	0.000
	Mix K-6	14	27	+0.10	0.02	5.19	13.07	0.000
Student Achievement Level	Low Achievers	33	48	+0.14	0.02	5.83	11.75	0.000
	Moderate/high Achievers	11	15	+0.07	0.03	2.41	12.06	0.033
	Mixed Achievers	59	114	+0.08	0.01	5.56	31.66	0.000
Socio-Economic Status	Low SES	31	50	+0.08	0.03	2.70	31.68	0.011
	Moderate/high SES	50	73	+0.11	0.02	6.05	33.36	0.000
	Mixed SES	25	54	+0.09	0.02	4.74	21.24	0.000
U.S. vs. Other Countries	U.S. Studies	63	126	+0.10	0.02	6.41	36.08	0.000
	Non-U.S. Studies	22	51	+0.07	0.03	2.78	19.73	0.012
Tutoring Group Size	One-to-one	8	13	+0.19	0.05	3.53	7.58	0.008
	One-to-small	13	22	+0.30	0.05	5.47	12.44	0.000
Tutoring Specific Moderators (Exploratory Only)								
Tutoring Provider	Teachers	3	6	+0.23				
	Teaching Assistants	17	28	+0.19				
	Paid Volunteers	1	1	+0.20				
Tutoring Group Size and Provider	One-to-One by Teachers	2	5	+0.22				
	One-to-One by Teaching Assistants	5	7	+0.18				
	One-to-One by Paid Volunteers	1	1	+0.20				
	One-to-Small Group by Teachers	1	1	+0.34				
	One-to-Small Group by Teaching Assistants	12	21	+0.30				

Note. k = number of studies; *n* = number of outcomes; ES=effect size; SE=standard error; df=degrees of freedom.

Exploratory model is the same as the full model, adding the tutoring provider moderator. Because of the limited sample size and exploratory nature, statistical tests are not reported.

Figure 1.
PRISMA Flow Diagram of Study Search and Review Process.



*A total of 82 unique citations were included in the review. Of those citations, some reported on more than one intervention, so they are included as having multiple studies, bringing the total number of included studies to 85.

Figure 2
Theories of Action Leading to
Categories of Improvement Strategies for Elementary Mathematics