Effective Programs in Elementary Mathematics: A Meta-Analysis

Marta Pellegrini (ID)

University of Florence

Cynthia Lake
Amanda Neitzel (ID)
Robert E. Slavin

Johns Hopkins University

This article reviews research on the achievement outcomes of elementary mathematics programs; 87 rigorous experimental studies evaluated 66 programs in grades K–5. Programs were organized in six categories. Particularly positive outcomes were found for tutoring programs (effect size [ES] = +0.20, k = 22). Positive outcomes were also seen in studies focused on professional development for classroom organization and management (e.g., cooperative learning; ES = +0.19, k = 7). 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 (nondigital) curricula (ES = +0.12, k = 7), but not for digital curricula. Traditional and digital curricula with limited professional development, as well as benchmark assessment programs, found few positive effects.

Keywords: *evidence of effectiveness*

In recent years, there has been an increasing emphasis on the identification and dissemination of programs proven in rigorous experiments. This emphasis has been clear in federal funding for education research, especially at the Institute for Educational Sciences (IES), Education Innovation Research (EIR), and the National Science Foundation (NSF). The establishment of the What Works Clearinghouse (WWC) has helped establish standards of evidence and has disseminated information on the evidence base for educational programs. In England, the Education Endowment Foundation has similarly supported rigorous research in education. In 2015, the Every Student Succeeds Act defined, for the first time, criteria for the effectiveness of educational programs. Every Student Succeeds Act (ESSA) places particular emphasis on three top levels of evidence: strong (statistically significant positive effects in at least one randomized experiment), moderate (statistically significant positive effects in at least one quasi-experiment), and promising (statistically significant positive effects in at least one correlational study). ESSA encourages use of programs meeting these criteria, and requires schools seeking school improvement funding to adopt programs meeting one of these criteria.

One of the subjects most affected by the evidence movement in education is mathematics, because there is more rigorous research in mathematics than in any other subject except reading. The rapid expansion in numbers and quality of studies of educational programs has provided a far stronger basis for evidence-informed practice in mathematics than once existed.

The advances in research have been noted in reviews, cited later in this article. However, the great majority of reviews have focused only on particular approaches or subpopulations, using diverse review methods. This makes it difficult to compare alternative approaches on a consistent basis, to understand the relative impacts of different programs. The most recent meta-analyses to systematically review research on all types of approaches to mathematics instruction were a review of elementary mathematics programs by Slavin and Lake (2008) and one by Jacobse and Harskamp (2011). A meta-analysis of all secondary mathematics programs was published by Slavin et al. (2009).

The present article updates the Slavin and Lake (2008) review of elementary mathematics, incorporating all rigorous evaluations of programs intended to improve mathematics achievement in grades K–5. The review uses more rigorous selection criteria than would have been possible in 2008, and uses current methods for meta-analysis and meta-regression, to compare individual programs and categories of programs, as well as key mediators, on a consistent basis.

Two reviews considering all elementary mathematics programs have been published since 2008. Slavin and Lake (2008) identified 87 qualifying studies of outcomes of elementary mathematics programs and concluded that mathematics programs that incorporate cooperative learning, classroom management, 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 40 studies. The authors reported that small group or individual interventions had greater effects on mathematics achievement than did whole-class programs.

An important contribution of the present review is its focus on coherent categories of mathematics interventions. Most previous reviews of mathematics interventions have focused on variables rather than programs or categories of similar programs (e.g., Gersten et al., 2014; Lynch et al., 2019). Yet to inform practice in elementary mathematics, it is important to identify specific effective programs and categories of programs, because this is how educators and policymakers interested in evidence-based reform make choices (Morrison et al., 2019). For example, the 2015 ESSA defines *program* effectiveness, and the WWC (2020) is similarly focused on evaluating evidence for *programs*, not variables.

The importance of program categories stems from the importance of programs. A daunting problem in evidence-based reform in education is that few programs are supported by large numbers of rigorous studies. The vast majority of practical programs with any rigorous evidence of effectiveness at all have just one or two studies that would meet modern standards. If there are several similar programs that also find positive impacts in rigorous experiments, this may buttress the claims of effectiveness for all of them. On the contrary, if a given program shows positive impacts in a single rigorous experiment, but other equally rigorous studies of similar programs do not, this should cause educators and researchers to place less confidence in the one study's findings.

In the present meta-analysis, we included all studies that met a stringent set of inclusion criteria, regardless of the type of program used. We then grouped the programs into six mutually exclusive categories. These are described in detail later in this article, but in brief, the categories are as follows:

1. Tutoring (e.g., one-to-one or one-to-small group instruction in mathematics)
2. Professional development (PD) focused on mathematics content and pedagogy (at least 2 days or 15 hours)
3. PD (at least 2 days or 15 hours) focused on classroom organization and management (e.g., cooperative learning in mathematics)
4. PD focused on implementation of traditional (non-digital) and digital curricula (at least 2 days or 15 hours)
5. Traditional and digital curricula with limited PD (less than 2 days or 15 hours)
6. Benchmark assessments

A major feature of the present review is its use of modern approaches to meta-analysis and meta-regression that enable researchers to control effects of programs, categories and variables for substantive and methodological factors, and to obtain meaningful estimates for key moderators (see Borenstein et al., 2009; Borenstein et al., 2017; Lipsey, 2019; Pigott & Polanin, 2020; Valentine et al., 2019).

Another important contribution of the present meta-analysis is its use of stringent inclusion standards, similar to those of the WWC (2020). For example, the review of research on elementary mathematics programs by Slavin and Lake (2008), mentioned earlier, required that studies use random assignment or quasi-experimental designs, excluded measures overaligned with the treatment, and required a minimum duration of 12 weeks and a minimum sample size of 30 students in each treatment group. This review found positive effects for PD approaches, such as cooperative learning, mastery learning, and classroom organization and management, which had a mean effect size (ES) of $+0.33$ ($k = 36$). Technology-focused programs had a mean ES of $+0.19$ ($k = 38$), and curriculum approaches (mostly textbooks) had a mean ES of $+0.10$ ($k = 13$). These ESs are in a range similar to those reported by WWC (2013) studies of K–12 mathematics. The Lynch et al. (2019) review used similar inclusion standards, and reported an overall impact on mathematics learning of $+0.27$. Yet other reviews of mathematics interventions find much larger overall impacts. This is due to their inclusion of studies with design features known to significantly inflate ESs. For example, the third meta-analysis to include all studies of elementary mathematics, Jacobse and Harskamp (2011), reported an average ES of $+0.58$, about twice the size of the Slavin and Lake (2008) and Lynch et al. (2019) mean ESs. They noted that the review studies using non-standardized measures obtained significantly larger ESs than those using standardized measures, yet they did not control for this difference, known from other research (e.g., Cheung & Slavin, 2016) to be a powerful methodological factor in achievement ESs.

In recent years, research has established the substantial inflationary bias in ES estimates introduced by certain research design elements. Particularly important sources of bias include small sample size, very brief duration, use of researchers rather than school staff to deliver experimental programs, and use of measures made by developers and researchers (Cheung & Slavin, 2016; de Boer et al., 2014; Wolf et al., 2020).

The problem is that despite convincing demonstrations of the biasing impact of these factors, most reviews of research

do not exclude or control for studies that contain factors known to substantially and spuriously inflate ESs. As a result, meta-analyses often report ESs that are implausibly large. As a point of reference, a study by Torgerson et al. (2013) found an ES of +0.33, the highest for one-to-one tutoring in mathematics by certified teachers in the current review. How could studies of far less intensive treatments produce much larger effects than one-to-one tutoring?

As one example, a review of research on intelligent tutoring systems by Kulik and Fletcher (2016), mostly in mathematics, reported an implausible ES of +0.66. The review had a minimum duration requirement of only 30 minutes. The review reported substantial impacts of "local" (presumably researcher-made) vs. standardized measures, with means of +0.73 and +0.13, respectively. It reported ESs of +0.78 for sample sizes less than 80, and +0.30 for sample sizes over 250. Individual included studies with very low sample sizes reported remarkable (and implausible) ESs. A 50-minute study involving 48 students had an ES on local measures of +0.95. Another, with 30 students and a duration of one hour, found an ES of +0.78. A third, with 30 students and a duration of 80 minutes, reported an ES of +1.17. Yet in its overall conclusions, Kulik and Fletcher (2016) did not exclude or control for inclusion of very small or very brief studies or inclusion of "locally developed" measures and did not weight for sample size. In a separate analysis, the review reported on 15 mostly large, long-term studies of a secondary technology program called Cognitive Tutor, showing ESs of +0.86 on "locally developed" measures and +0.16 on standardized measures, but simply averaged these to report an ES of +0.45, an implausibly large impact. As a point of comparison, the WWC, which uses inclusion criteria similar to those used by Slavin and Lake (2008) and Lynch et al. (2019), accepted five studies of Cognitive Tutor Algebra I, which had a median ES of +0.08, and one of Cognitive Tutor Geometry with an ES of −0.19.

As another example, Lein et al. (2020), in a review of research on word problem solving interventions, reported mean ESs of +0.68 for researcher-made measures, compared with +0.09 for norm-referenced measures. They also reported a mean of +0.71 for interventions delivered by researchers, compared with +0.28 for those delivered by school staff. Yet the review did not control for these or other likely biasing factors and reported an implausible mean ES of +0.56.

In the present meta-analysis, we used inclusion criteria more stringent than those used by the WWC or by Slavin and Lake (2008) or Lynch et al. (2019), and substantially more stringent than those of the great majority of reviews of studies of mathematics programs. We excluded all measures made by developers or researchers, post hoc quasi-experiments, very small and very brief studies, and those in which researchers, rather than staff unaffiliated with the research taught the experimental program. We also weighted studies

by their sample sizes (using inverse variance) in computing mean ESs. Then we statistically controlled for relevant methodological and substantive moderators. These methods are described later in this article.

The importance of these procedures should be clear. Whatever outcomes are reported for studies included in the present meta-analysis, readers should be able to be confident that these outcomes are due to the actual likely effectiveness of the interventions, not to methodological or substantive factors that are known to bias ES estimates from extensive prior research. Failing to exclude or control for these factors not only spuriously inflates reported ESs but it also confounds comparisons of ESs within reviews, as a program's large ES could be due to use of study features known to inflate ESs in the studies evaluating it, rather than to any actual greater benefit for students.

The inclusion of studies with certain study features not only risks substantial inflation of mean ESs, but also may undermine the relevance of the study for practice. A study of 30 minutes' duration, one that has a sample size of 14, one that uses researchers rather than school staff to deliver the intervention, or one that uses outcome measures created by developers or researchers, is of little value to teachers or students, because educators need information on what works over significant time periods, is implemented by school staff, and is evaluated using universally accepted assessments, not ones they themselves made up.

Method

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 similar to those of the WWC (2020), with a few exceptions noted below. A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 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 intended to improve mathematics achievement in 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
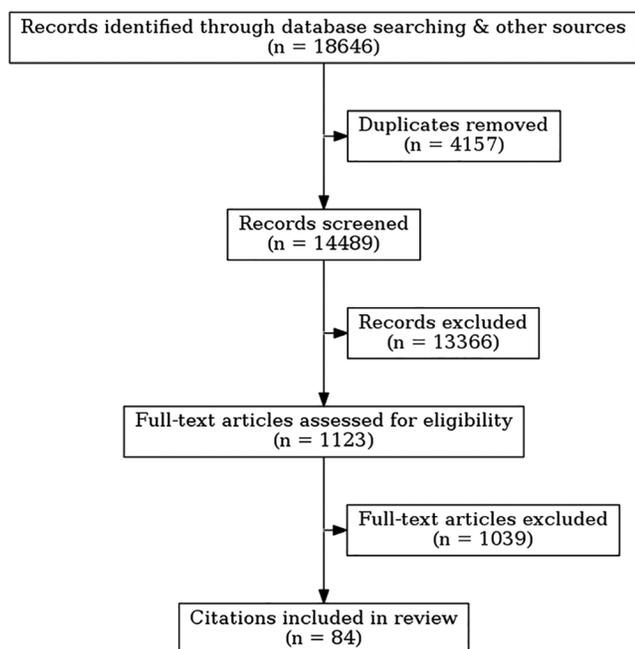
FIGURE 1.   *PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study search and review process.*

Note. A total of 84 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 87.

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 school staff unaffiliated with the research, not by the program developers, researchers, or their graduate students. This is particularly important for relevance to practice.

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. The WWC (2020) excludes "overaligned" measures, but not 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 ESs of independent measures on average, even within the same studies (Cheung & Slavin, 2016; de Boer et al., 2014; Gersten et al., 2009; Kulik & Fletcher, 2016; Lein et al. 2020; Lynch et al., 2019; Nelson & McMaster, 2019). Results from developer- or researcher-made measures may be valuable to researchers or theorists, and there are situations in which independent measures do not exist. However, such findings should only be supplemental information, not reported as outcomes of the practical impact of treatments.

8. Studies had to have a minimum duration of 12 weeks, to establish that effective programs could be replicated over extended periods. Also, very brief studies have been found to inflate ESs (e.g., Gersten et al., 2014; Kulik & Fletcher, 2016; Nelson & McMaster, 2019).

9. Studies could have taken place in the United States or in similar countries: Europe, Israel, Australia, or New Zealand. However, the report had to be available in English. In practice, all qualifying studies took place in the United States, the United Kingdom, Canada, the Netherlands, and Germany.

10. Studies had to have been carried out from 1990 through 2020, 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 multistep 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 education 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 reviewed studies accepted by the WWC, and searched in recent tables of contents of eight key mathematics and general educational journals from 2013 to 2020: *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 et al., 2017; Gersten et al., 2014; Jacobse & Harskamp, 2011; Kulik & Fletcher, 2016; Li & Ma, 2010; Lynch et al., 2019; Nelson & McMaster, 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 et al., 2016). 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 authors of the current study. 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.

Initial searching identified 18,646 potential studies. After removing 4,157 duplicate records, these search strategies yielded 14,489 studies for screening. The screening phase eliminated 13,366 studies, leaving 1,123 full-text articles to be assessed for eligibility. Of these full-text articles that were reviewed, 1,039 studies did not meet the inclusion criteria, leaving 84 contributions included in this review, with two studies including multiple interventions, for a total number of 87 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 ESs.

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), socioeconomic status (low SES vs. moderate/high SES), and study locations in the United States. versus 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). The coded data are available on GitHub (Pellegrini et al., 2021).

Effect Size Calculations and Statistical Procedures

ESs 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 ESs when unadjusted standard deviations were not available.

Statistical significance is reported for each study using procedures from the 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, 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 (Raudenbush & Bryk, 2002). Studies with cluster assignments that did not use hierarchical linear modeling or other multilevel 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 ESs 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 ESs per study, and RVE 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 ESs (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 ES 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 ESs 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. 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. The theoretical base for tutoring draws on research in reading, which has long made extensive use of one-to-one and small group tutoring (see, e.g., Elbaum et al., 2000; Gersten et al., 2020; Slavin et al., 2011; Wanzek et al., 2016) as well as in mathematics (e.g., Fuchs, Schumacher, et al., 2013; Fuchs, Schumacher, et al., 2016; Jacobse & Harskamp, 2011; Nelson & McMaster, 2019). 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 substantially adapt their instruction to the needs of the student(s). Tutoring programs in mathematics generally provide well-structured, sequential materials for students, but tutors are trained to explain and demonstrate concepts for students who are struggling with it. Tutors are trained to start with struggling students where they are and move them forward rapidly. They are able to explain and model mathematical concepts and processes, observe how students are working, and give them personalized feedback and encouragement. Tutors can enable students to work in small steps, experiencing success at each step. Furthermore, tutors are likely to be able to build close personal relationships with tutored student(s), giving them attention and praise that many students crave, and enhancing their motivation as students seek to please a valued adult. Previous reviews of research on elementary mathematics approaches have found that tutoring is among the most effective of all interventions for students struggling in mathematics (e.g., Jacobse & Harskamp, 2011; Slavin & Lake, 2008).

2. *PD Focused on Mathematics Content and Pedagogy*. Interventions in this category provide intensive content-focused PD intended to advance teachers' understanding of current standards-based content and effective pedagogy (teaching methods). To be included in this category, PD had to be provided for at least 2 days or 15 hours. This category of strategies emphasizes giving teachers knowledge about mathematics content and about ways of explaining it (Desimone, 2009; Desimone & Garet, 2015; Kennedy, 2014; Penuel et al., 2011). Ideally, such approaches emphasize mathematics *content, active learning, coherence, sustained duration*, and *collective participation* to help teachers learn and apply to their teaching new understandings of mathematics content and mathematics-specific pedagogy (Desimone, 2009; Desimone & Garet, 2015; Kennedy, 2014; Penuel et al., 2011). Almost all of these PD programs (as well as those in Categories 3, 4, and 5) provided some degree of on-site coaching to follow up after initial training. Coaching has been found to be an effective component of PD in mathematics (Kraft et al., 2018).

3. *PD Focused on Classroom Organization and Management*. This mathematics-specific category includes programs that provide teachers with PD and materials to help them implement innovations in classroom organization and management, such as cooperative learning (e.g., Slavin, 2017) and classwide behavior approaches (e.g., Weis et al., 2015). This category had the highest ES (*ES* = +0.33, *k* = 36) of any category in the Slavin and Lake (2008) meta-analysis. Previous research on cooperative learning has shown positive effects on mathematics and other subjects (e.g., Rohrbeck et al., 2003; Webb, 2008).

4. *PD Focused on Implementation of Traditional and Digital Curricula*. Interventions in this category provide teachers with moderate to extensive PD (at least 2 days or 15 hours, combining training and follow-up coaching) to support informed, thoughtful implementation of innovative traditional (i.e., non-digital) or digital curricula for students. There were two subcategories: (a) PD Focused on Implementation of Traditional Curricula, with minimal use of technology and (b) PD Focused on Implementation of Digital Curricula, such as computer-assisted instruction.

TABLE 1
Meta-Regression Results

Coefficient	Reference group	β	*SE*	*t*	*df*	*p*
Null model						
Intercept		0.11	0.02	6.42	72.92	.000
Meta-regression						
Intercept	Tutoring	0.10	0.01	7.93	41.95	.000
PD focused on classroom organization and management		0.04	0.08	0.48	8.89	.641
PD focused on mathematics content and pedagogy		−0.12	0.07	−1.75	23.55	.094
PD focused on implementation of traditional and digital curricula		−0.15	0.07	−2.26	10.26	.047
Traditional and digital curricula with limited professional development		−0.11	0.07	−1.63	17.85	.120
Benchmark assessments		−0.15	0.10	−1.56	7.10	.163
PD focused on implementation of traditional curricula	PD focused on implementation of digital curricula	0.12	0.04	2.78	7.33	.026
Digital curricula	Traditional curricula	0.04	0.04	1.01	24.56	.324
Quasi-experiments	Randomized studies	0.12	0.04	3.30	12.21	.006
K–2	Mixed	−0.04	0.03	−1.15	15.79	.267
3–6		0.00	0.02	−0.09	11.51	.930
Low achievers	Mixed achievers	0.05	0.03	1.87	12.07	.086
Moderate/high achievers		−0.02	0.02	−0.84	12.10	.419
Low SES	Mixed SES	−0.02	0.02	−0.65	20.74	.524
Moderate/high SES		0.01	0.02	0.31	22.36	.759
International studies	U.S. Studies	−0.02	0.03	−0.47	30.80	.643
One-to-small group tutoring	One-to-one tutoring	0.12	0.08	1.52	15.19	.149

Note. Meta-regression model also controlled for cross-age and online tutoring. PD = professional development; SES = socioeconomic status.

5. *Traditional and Digital Curricula With Limited PD* includes two subcategories: (a) Traditional (i.e., nondigital) curricula (textbooks with associated teaching materials) and (b) Digital curricula for students. Limited PD (less than 2 days or 15 hours) was provided in such strategies.

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., Konstantopoulos et al., 2016).

Results

A total of 87 studies evaluating 66 programs met the inclusion standards of this review. The studies included were of high methodological quality: 74 (85%) of the studies were randomized trials and 13 (15%) were quasi-experimental studies. Also, 75 (86%) of the studies were reported in 2010 or later, indicating the extraordinary pace at which rigorous studies of elementary mathematics are appearing. Only four of the studies included in the current review overlapped those cited by Slavin and Lake (2008). Studies cited in 2008

but not in the current article were released before 1990, or did not meet the much more stringent inclusion requirements of the current synthesis.

Table 1 shows the meta-regression outcomes. The full model controlled for program category and subcategory, research design, grade level, student achievement level, SES, the United States versus 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 ES of +0.09, $p < .01$ ($k = 87$), with outcomes that vary substantially among different categories.

Tutoring Programs

Twenty-two studies evaluated tutoring programs. Combining all forms of tutoring, the mean ES was +0.20, $p < .01$ ($k = 22$). 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 United Kingdom, and another

TABLE 2

Mean Effect Sizes of Program Categories and Subcategories

Table	Category	k	n	ES	SE	t	df	p
3	Tutoring programs	22	39	+0.20	0.05	4.21	7.86	.003
	One-to-one tutoring	8	13	+0.19	0.06	3.36	7.50	.011
	One-to-small group tutoring	14	26	+0.30	0.05	5.88	13.38	.000
4	Professional development focused on mathematics content and pedagogy	10	23	+0.03	0.03	0.86	9.01	.411
5	Professional development focused on classroom organization and management	7	11	+0.19	0.06	3.30	4.16	.028
6	Professional development focused on implementation of traditional and digital curricula	12	35	+0.01	0.03	0.42	3.13	.705
	Professional development focused on implementation of traditional curricula	7	18	+0.12	0.03	4.88	5.51	.003
	Professional development focused on implementation of digital curricula	5	17	0.00	0.03	−0.03	3.15	.977
7	Traditional and digital curricula with limited professional development	30	67	+0.05	0.03	1.52	12.52	.153
	Traditional curricula	15	34	+0.04	0.04	1.06	12.33	.309
	Digital curricula	15	33	+0.08	0.02	4.02	11.88	.002
8	Benchmark assessments	4	5	0.00	0.08	−0.03	3.12	.975

Note. k = number of studies; *n =* number of outcomes; *ES =* effect size.

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. Fourteen studies evaluated programs taught by tutors to small groups. Overall, the weighted mean ES 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 ES of −0.03 and the one study of cross-age peer tutoring had an ES of +0.02. One-to-one tutoring by certified teachers (ES = +0.22, $k = 2$), and by teaching assistants (*ES* = +0.16, $k = 5$) were not significantly different from each other 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 ES of +0.30, $p < .01$, $k = 14$). Surprisingly, outcomes of one-to-small group tutoring using structured programs were (nonsignificantly) higher than those of one-to-one tutoring. The only one-to-small group program that used certified teachers (ES = +0.36, $k = 1$) was similar in outcomes to one-to-small group approaches that used teaching assistants as tutors (ES = +0.30, $p < .01$, $k = 13$). 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

Ten studies evaluated 10 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 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 ES was +0.03, *ns* ($k = 10$) for all professional development programs focused on mathematics content and pedagogy.

Professional Development Focused on Classroom Organization and Management

Professional development approaches in this category focused on helping teachers use models such as cooperative learning and classroom management strategies (see Table 5). Across seven studies of six diverse programs, the average ES for mathematics was +0.19, $p < .01$ ($k = 7$).

Professional Development Focused on Implementation of Traditional (Nondigital) and Digital Curricula

Twelve 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 ES was +0.01, *ns* ($k = 12$). ESs

TABLE 3
Tutoring Programs

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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 WJ-App. Problems WJ-Quant Concepts WJ-Math Reasoning	+0.15* +0.28* +0.24* +0.30*	+0.24*
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						Program mean: +0.05		
Catch Up® Numeracy								
Hodgen et al. (2019)	CR	1 Year	142 Schools, 1,481 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, Geary, et al. (2013)	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*

9

TABLE 3 (CONTINUED)

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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, 1,201 students (567E, 634C)	Year 6 (Grade 5)	England. 48% FRL, 72%W	Key Stage 2		+0.19
FocusMATH								
Styers and 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, Schumacher, et al. (2016)	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, Malone, et al. (2016)	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, 1,089 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 NSB SESAT	+0.32* +0.16 +0.001	+0.16

(continued)

TABLE 3 (CONTINUED)

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
Doabler et al. (2016)	SR	5 Months	292 Students (208E, 82C)	K	Boston. 7% AA, 89% W, 50% H, 26% ELL.	TEMA–3 NSB SESAT	+0.31* +0.40* +0.24	+0.32*
Clarke et al. (2017)	SR	4 Months	689 Students (527E, 162C)	K	Oregon. 55% W, 26% H, 26% ELL, 87% FRL.	TEMA–3 NSB SESAT	+0.25* +0.09 +0.12	+0.15
Working Memory Intervention								
Wright et al. (2019)	CR	5 Months	171 Schools, 1,822 students (882E, 940C)	Year 3 (Grade 2)	England; 37% FRL, 80% W	GL Assessment British Ability		+0.22
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); 2,786 students; Year 5 (tutors); 2,683 students	Year 3, 5 (Grades 2, 4)	England; 22% FRL, 86% W, 4% AA, 5% A	ICAS-Year 3 ICAS-Year 5	+0.01 +0.02	+0.02

Note. 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 State 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-price 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 effect size
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. (2020)	CR	1 Year	22 Schools, 2,005 students (1,046E, 959C)	1, 2	FL; 60% FRL, 18% AA, 37% H, 37% W, 23% ELL	ITBS; Grade 1 Comp. / Grade 1 Problems / Grade 2 Comp. / Grade 2 Problems	+0.03 / +0.14 / −0.29 / −0.07	−0.04
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 / NWEA	−0.06* / −0.05	−0.05*
Math for All								
Duncan et al. (2018)	CR	1 Year	29 Schools, 881 students (423E, 458C)	4, 5	Chicago 79% FRL, 37% AA, 42% H, 23% ELL	NWEA MAP		+0.11
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 / Grade 5	+0.04 / +0.08	+0.06
PBS TeacherLine								
Dominguez et al. (2006)	CR	1 Year	87 Teachers, 1,119 students (523E, 596C)	3–5	FL, SC, NY	Algebra test / Geometry test	−0.02 / +0.08	+0.03
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 the 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

Note. 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 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-price lunch; ELL = English language learner; LD = learning disabilities; SPED = special education.

*p < .05 at the appropriate level of analysis (cluster or individual).

TABLE 5
Professional Development Focused on Classroom Organization and Management

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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								
VanDerHeyden 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
TAI							Program mean: +0.11	
Stevens and 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 and 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	
Math PALS								
Wood et al. (2020)	CR	1 Year	28 Teachers, 454 students (205E, 249C)	1	FL; 45% FRL, 84% W, 4% AA, 3% H	WJ-III Math Fluency	+0.16	+0.11
						WJ-III Applied Problems	+0.06	

Note. 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-price lunch; ELL = English language learner; LD = learning disabilities; SPED = special education.
*p < .05 at the appropriate level of analysis (cluster or individual).

13

TABLE 6

Professional Development Focused on Implementation of Traditional (Nondigital) and Digital Curricula

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
Category mean: +0.01								
Professional development focused on implementation of traditional (nondigital) 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 / Program mean: +0.10	Number Knowledge Test		+0.10
Mathematics Reasoning								
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*

(continued)

TABLE 6 (CONTINUED)

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
Professional development focused on implementation of digital curricula								
Subcategory mean: 0.00								
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, 921 students (941E, 980C)	5	Urban, rural and suburban schools in West Virginia. 94% W, 50% FRL	WVGSA		−0.13
Wang and Woodworth (2011a)	SR	4 Months	651 Students (521E, 130C)	2–5	San Francisco Bay Area; 7% H, 81% ELL, 88% FRL	NWEA-Math Over.	−0.02	−0.01
						NWEA-Probl. Solv.	−0.05	
						NWEA-Num. sense	+0.01	
						NWEA-Comp.	−0.08	
						NWEA-Geometry	+0.11	
						NWEA-Statistics	−0.02	
Time to Know								
Rosen and Beck-Hill (2012)	CQE	6 months	4 Schools, 476 students (283E, 193C)	4–5	Dallas, TX; 18% AA, 63% H	TAKS		+0.31

Note. 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 Test; NSB = 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-price lunch; ELL = English language learner; LD = learning disabilities; SPED = special education.

*p < .05 at the appropriate level of analysis (cluster or individual).

15

TABLE 7

Traditional (Nondigital) and Digital Curricula With Limited Professional Development

Category mean: +0.05
Traditional (nondigital) curricula
Subcategory mean: +0.04
Early Learning in Mathematics

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
Resendez and 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 and 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. / TerraNova-Comp. / TerraNova-Comp.	+0.10 / −0.21 / +0.05	−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. / MAT-Math Comp. / GMADE	−0.13 / +0.06 / −0.06	−0.04
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

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
Vaden-Kiernan et al. (2015)	CR	2 Years	48 Schools, 4,467 students	K–5	51% AA, 73% FRL.	GMADE		−0.01

GO Math!

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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, Grade 1 / Grade 2	0.00 / +0.09	+0.04
Gatti and Giordano (2008)	CR	1 Year	77 Classes, 1,363 students (729E, 634C)	1, 4	AZ, MA, OR, SC; 52% FRL, 27% H, 9% AA.	GMADE, Grade 1 / Grade 4	−0.14 / −0.31	−0.22*

JUMP Math

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
Solomon et al. (2011)	CR	5 Months	18 Schools, 267 students (163E, 104C)	5	Rural Canadian schools, Ontario.	WJ-III		+0.23

(continued)

TABLE 7 (CONTINUED)

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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*			
Educational Research Institute of America (2010)	QE	1 Year	678 Students (125E, 553C)	4	NJ. 15% FRL, 30% minority, 12% SPED	NJ ASK		+0.25*
Educational Research Institute of America (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: +0.08* **Accelerated Math**					Program mean: +0.02			
Lambert et al. (2014)	CR	1 Year	36 Classes, 504 students (256E, 248C)	2–5	Midwestern United States; 40% minority, 76% FRL, 18% SPED	TerraNova		+0.02
Lehmann and 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 and 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, 2,133 students (1103E, 1030C)	Year 4, 5 (Grade 3, 4)	England; 30% FRL	ACERs Essential Learning Metric (ELM)		−0.04

(continued)

TABLE 7 (CONTINUED)

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
DreamBox Learning					Program mean: +0.10			
Lenard and 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) North Carolina End-of-Grade EOG (3–5)	+0.12* +0.03	+0.08
Wang and Woodworth (2011b)	SR	4 Months	557 Students (446E, 111C)	K, 1	San Francisco Area; 87% H, 81% ELL, 88% FRL	NWEA-Math Over. NWEA-Probl. Solv. NWEA-Num. sense NWEA-Comp. NWEA-Geometry NWEA-Statistics	+0.11 +0.06 +0.08 +0.13 +0.16* +0.12	+0.11
Educational Program for Gifted Youth (EGPY)								
Suppes et al. (2013)	SR	1 Year	1,484 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 Year: 34 schools, 10,455 students; 2 years: 18 schools, 2,677 students	3–5	Southern CA; 90% FRL, 85% H, 63% ELL	CST 1 Year 2 Years	+0.09 +0.03	+0.08
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 Grade 3 Grade 5	+0.11 +0.03	+0.07
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 and 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 GMADE-Grade 5	+0.27 −0.19	+0.06

(continued)

TABLE 7 (CONTINUED)

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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
Stop and Think								
Roy et al. (2019)	CR	1 Year	84 Year groups, 2,702 students (1343E, 1359C)	Year 3, 5 (Grade 2, 4)	England; 30% FRL	Progress Test in Maths (PTM)		+0.09

Note. 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; WI III = Woodcock-Johnson III. Demographics: A = Asian; AA = African American; H = Hispanic; W = White; FRL = free/reduced-price lunch; ELL = English language learner; LD = learning disabilities; SPED = special education.

*p < .05 at the appropriate level of analysis (cluster or individual).

TABLE 8
Benchmark Assessments

Study	Design	Duration	Sample size	Grade	Sample characteristics	Posttest	Effect size	Study effect size
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, 13,944 students (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*

Note. 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-price lunch; ELL = English language learner; LD = learning disabilities; SPED = special education.
*$p < .05$ at the appropriate level of analysis (cluster or individual).

averaged $+0.12$, $p < .01$ ($k = 7$) for traditional (nondigital) curricula, but 0.00, *ns* ($k = 5$) for digital curricula.

Traditional and Digital Curricula With Limited Professional Development

Thirty studies evaluated 19 mathematics curricula, primarily traditional (nondigital) 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 ES was $+0.05$, *ns* ($k = 30$). Fifteen studies of traditional curricula, mostly textbooks, found a mean ES of $+0.04$, *ns* ($k = 15$), and 15 studies that evaluated digital curricula found a mean ES of $+0.08$, $p < .01$ ($k = 15$).

Benchmark Assessments

Four studies evaluated three programs that use benchmark assessments, summarized in Table 8. The studies found a mean ES of 0.00, *ns* ($k = 4$).

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, ESs may vary according to research design. Cheung and Slavin (2016) and de Boer et al. (2014) found that quasi-experiments across all subjects and grade levels, pre-K–12, produce a significantly higher ES than randomized studies, on average, although others, such as Lipsey and Wilson (2001), have not found this difference. In the present meta-analysis, differences in ESs between studies that used randomized designs (ES $= +0.08$, $p < .01$, $k = 74$) and studies that used quasi-experimental designs incorporating matching (ES $= +0.20$, $p < .01$, $k = 13$) were tested. This difference ($\beta = 0.12$) was significant ($p < .05$).

TABLE 9
Methodological and Substantive Moderators

Moderator	Level	k	n	ES	SE	t	df	p
Research design	Quasi-experiments	13	20	+0.20	0.03	6.16	9.77	.000
	Randomized studies	74	164	+0.08	0.01	6.33	39.14	.000
Grade level	K–2	33	79	+0.08	0.02	3.92	26.77	.001
	3–6	46	78	+0.11	0.02	4.38	34.78	.000
	Mix K–6	14	27	+0.11	0.02	6.29	10.85	.000
Student achievement level	Low achievers	33	48	+0.13	0.02	5.58	13.61	.000
	Moderate/high Achievers	11	15	+0.06	0.03	2.43	12.00	.032
	Mixed Achievers	61	121	+0.08	0.01	6.43	33.71	.000
Socioeconomic status	Low SES	36	56	+0.08	0.02	3.31	33.59	.002
	Moderate/high SES	52	73	+0.10	0.02	6.54	34.80	.000
	Mixed SES	26	55	+0.10	0.02	6.29	22.55	.000
USA vs. other countries	U.S. Studies	65	136	+0.10	0.02	6.58	36.94	.000
	Non-U.S. Studies	22	48	+0.08	0.03	3.07	20.95	.006
Tutoring Group Size	One-to-one	8	13	+0.19	0.06	3.36	7.50	.011
	One-to-small group	14	26	+0.30	0.05	5.88	13.38	.000
Tutoring Specific Moderators (Exploratory Only)								
Tutoring provider	Teachers	3	6	+0.24				
	Teaching assistants	18	32	+0.18				
Tutoring group size and provider	One-to-one by teachers	2	5	+0.22				
	One-to-one by teaching assistants	5	7	+0.16				
	One-to-small group by teachers	1	1	+0.36				
	One-to-small group by teaching assistants	13	25	+0.30				

Note. k = number of studies; n = number of outcomes; ES = effect size. 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.

Grade Levels. To determine if different grade levels may be a source of variation, we divided the study outcomes into those relating to Grades K to 2 and those relating to grades 3 to 6. Many studies crossed this divide, so one study could contribute both a K–2 and a 3–5 ES. The mean ES for K–2 outcomes (ES = +0.08, $p < .01$, n = 79) was very similar to the mean ES for 3 to 6 outcomes (ES = +0.11, $p < .01$, n = 78).

Student Achievement Level. Outcomes including all students had a mean ES of +0.08, $p < .01$ (n = 121). This was not significantly different from either outcomes for low achievers (ES = +0.13, $p < .01$, n = 48) or outcomes for moderate and high achievers (ES = +0.06, $p < .05$, n = 15).

Socioeconomic 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 United States, 21% for England). Mean ESs for outcomes of mixed SES populations were +0.10, $p < .01$ (n = 55). The mean ES for low SES students was +0.08, $p < .05$ (n = 56), and for moderate/high SES students, it was +.10, $p < .01$ (n = 73). The differences

between mixed and low-SES students ($\beta = -0.01$, n.s.) and mixed and moderate/high-SES students ($\beta = 0.01$, *ns*) were not statistically significant.

The United States Versus Other Countries. Of the 87 qualifying studies, 65 took place in the United States, 19 in England, 1 in the the Netherlands, one in Germany, and one in Canada. Mean ESs were nearly identical for U.S. and non-U.S. studies: +0.10, $p < .01$ for U.S. (k = 65), +0.08, $p < .01$ for non U.S. (k = 22). This difference ($\beta = -0.02$, *ns*) 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 with those for tutoring provided in small-group settings (ES = +0.30, $p < .01$, k = 13). Outcomes were not significantly different ($\beta = 0.11$, *ns*).

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 ESs for five

21

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 = +0.24, *k* = 3) was similar to those of tutoring provided by teaching assistants (ES = +0.18, *k* = 18).

Discussion

This review of evaluations of elementary mathematics programs found 87 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. Strict inclusion criteria, plus controls for key moderators, made ES estimates much lower than those in previous meta-analyses, but because of these procedures, the ESs are more realistic than were those in studies with less strict inclusion standards (e.g., Jacobse & Harskamp, 2011; Kulik & Fletcher, 2016; Slavin & Lake, 2008). 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–12, so this is not a head-to-head comparison. But the relative outcomes are nevertheless interesting.

Both Lynch et al. (2019) and the present study found small, nonsignificant impacts for professional development services without a strong link to new curriculum, and both found small, nonsignificant impacts of implementation of traditional or digital curricula with a limited focus on professional development (less than 2 days or 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 (nondigital) curricula (ES = +0.12, *p* < .01), but found an ES 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 organization and management (ES = +0.19, *p* < .01, *k* = 7). Forms of cooperative learning were most common among such studies. The Lynch et al. (2019) meta-analysis did not identify a comparable category of professional development 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. 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 ES for one-to-small group tutoring (ES = +0.30, *p* < .01, *k* = 14) was larger than that for one-to-one (ES = +0.19, *p* < .01, *k* = 8), though this difference was not statistically significant. Similar findings were reported by Clarke et al. (2017). Teachers (ES = +0.24, *k* = 3) and teaching assistants (ES = +0.18, *k* = 18) appear equally effective as tutors, on average, but this result should be interpreted cautiously due to the small sample of teacher–tutor studies. In contrast, online 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, *p* < .01, *k* = 13) suggests that this tutoring arrangement 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 substantially adapt to the learning needs of students (e.g., Elbaum, 2000). Yet digital curricula also emphasizes individualization, and ESs for all studies using digital curricula had ESs near zero (see Tables 6 and 7). However, the kind of individualization possible in one-to-one or one-to-small group tutoring is beyond what technology can provide. As in CAI, successful tutoring programs in mathematics generally provide structured, sequential content at students' individual levels, and allow students to proceed at their own pace. However, face-to-face tutors can also individualize by providing feedback, explanations, and demonstrations to help students understand key concepts and get past blockages and misconceptions. Also, it may be that tutoring, by providing struggling students with individual attention from caring tutors, may provide a motivational or social–emotional benefit that computers cannot, and students may be eager to please a valued adult. Research is needed to understand the effectiveness of tutoring, including qualitative and correlational as well as experimental methods.

The positive effects of professional development focused on classroom organization and management (ES = +0.19, *k* = 7) replicate findings from previous reviews, such as Slavin and Lake (2008) and Jacobse and Harskamp (2011), as well as a great deal of research on cooperative learning (e.g., Rohrbeck et al., 2003; Slavin, 2017; Webb, 2008).

Programs in the classroom organization and management category generally assign students to teams and encourage them to help one another learn and behave appropriately. Teams receive recognition or small privileges (such as lining up first for recess) if their members, on average, behaved appropriately and performed well on assessments. Professional development is focused on facilitating teamwork, mutual assistance, encouragement, and commitment to

prosocial goals. All programs also focus on student success in mathematics. The success of this category of programs suggests that mathematics achievement may best be facilitated by enhancing motivation and making students active learners.

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 1–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 nonsignificant negative effects on Northwest Evaluation Association 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 10 studies of professional development methods focused on mathematics content and pedagogy achieved statistical significance in improving mathematics outcomes, and the mean was only +0.03. 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 experimental and 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 = +0.05, *ns, k* = 30). 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 mean ES of 0.00 (*k* = 4, *ns*).

One interesting finding from the present review relates to technology in mathematics education, which has been reviewed previously by Cheung and Slavin (2016); Higgins et al. (2019); Li and Ma (2010); and Savelsbergh et al. (2016). 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 (nondigital) curricula, essentially textbooks, had a modest positive impact on mathematics achievement, averaging ES = +0.12 (*k* = 7; *p* < .05). However, professional development supporting programs with a strong focus on technology had an average ES of 0.00. Among programs with limited professional development, both traditional curricula (*ES* = +0.04) and digital curricula (*ES* = +0.08) had minimal ESs (see Table 7). 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.

One might ask whether the impacts of digital curricula are increasing over time. To test this, we computed mean unweighted ESs for studies reported in three 5–year periods. The results were as follows:

Period	No. of studies (*k*)	Mean effect size
2005–2009	4	+0.04
2010–2014	9	+0.07
2015–2019	7	+0.05

Clearly, efforts of digital curricula are not improving. Across the 20 studies of digital curricula, there were two with impressive, though not statistically significant impacts: One was a study of Time to Know by Rosen and Beck-Hill (2012), with an ES of +0.31 (n.s.). The other was a study of Symphony Math, by Schwartz (2020), with an ES of +0.30 (*ns*). These may indicate promise for the next generation of digital curriculum, but they are single studies with too few schools to achieve adequate power for statistical significance (meaning that these large impacts could be due to characteristics of a few schools rather than true effects of the programs).

Technology in education has long been expected to have revolutionary impacts on learning. Computer-assisted instruction was expected to be effective because it places students at their precise level of proficiency, so that they need not repeat content they already know, and it advances students at their own pace, so that they can never fall behind. The computer, it is said, is patient, giving students as much time as they need to master the content, but moving forward rapidly if they are succeeding. The computer immediately provides answers, so students need not practice errors, but can correct themselves and move on. Every one of these points was made in a 1954 film that still exists on YouTube ("Teaching Machine and Programmed Learning"). Yet 67 years later, it is clear that technology programs based on these arguments, no matter how sensible they sound, have not transformed the outcomes of learning, not even in elementary mathematics (also not in elementary reading; see Neitzel et al., in press).

Across all approaches, effects were non-significantly larger for low achievers (ES = +0.13) than for others (moderate/high achievers: ES = +0.06, mixed achievers:

ES = +0.08), suggesting that there may be many pragmatic methods of increasing means while narrowing gaps.

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 and classroom management. 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.

Declaration of Conflicting Interests

Robert Slavin is a cofounder of TAI, a program named in this article.

ORCID iDs

Marta Pellegrini ![iD] https://orcid.org/0000-0002-9806-3231
Amanda Neitzel ![iD] https://orcid.org/0000-0002-4676-9320

Note

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 educational benefits.

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). U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED512551.pdf

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

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*. 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*. CNA Corporation. https://files.eric.ed.gov/fulltext/ED555557.pdf

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

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

*Clarke, B., Doabler, C. T., Smolkowski, K., Baker, S. K., Fien, H., & Strand Cary, M. (2016). Examining the efficacy of a Tier 2 kindergarten intervention. *Journal of Learning Disabilities*, *49*(2), 152–165. https://doi.org/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. https://doi.org/10.1080/02796015.2014.12087442

*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*(February), 97–113. https://doi.org/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. https://doi.org/10.3102/0034654314540006

Desimone, L. M. (2009). Improving impact studies of teacher's professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. https://doi.org/10.3102/0013189X08331140

Desimone, L. M., & Garet, M. S. (2015). Best practices in teachers' professional development in the United States. *Psychology, Society, & Education*, *7*(3), 252–263. https://doi.org/10.25115/psye.v7i3.515

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. https://doi.org/10.3102/0034654316687036

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

*Dominguez, P. S., Nicholls, C., & Storandt, B. (2006). *Experimental methods and results in a study of PBS TeacherLine Math Courses*. Hezel Associates. https://files.eric.ed.gov/fulltext/ED510045.pdf

*Duncan, T., Moeller, B., Schoeneberger, J., & Hitchcock, J. (2018). *Assessing the impact of the Math for All professional development program on elementary school teachers and their students*. http://mathforall.cct.edc.org/resources/

*Eddy, R. M., Hankel, N., Hunt, A., Goldman, A., & Murphy, K. (2014). *Houghton Mifflin Harcourt GO Math! Efficacy study year two final report*. Cobblestone Applied Research & Evaluation. https://s3.amazonaws.com/prod-hmhco-vmg-craftcms-public/research/HMH_GoMath_RCT_1_3_Final_Report_2014.pdf

*Educational Research Institute of America. (2010). *A study of the Singapore Math Program, Math in Focus, state test results* (Report # 404). Houghton Mifflin Harcourt. https://docplayer.net/21444009-A-study-of-the-singapore-math-program-math-in-focus-state-test-results.html

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

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, *92*(4), 605–619. https://doi.org/10.1037/0022-0663.92.4.605

*Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., DeSelms, J., Seethaler, P. M., Wilson, J., Craddock, C. F., & Bryant, J. D. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology*, *105*(1), 58–77. https://doi.org/10.1037/a0030127

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

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

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

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

*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). U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED569154.pdf

*Gatti, G. G. (2009). *Pearson SuccessMaker math pilot study. 2008-09 Final report*. Gatti Evaluation.

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

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

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

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools: A practice guide (NCEE 2009–4060)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., & Jayanthi, M. (2020). Meta-analysis of the impact of reading interventions for students in the primary grades. *Journal of Research on Educational Effectiveness*, *13*(2), 401–427. https://doi.org/10.1080/19345747.2019.1689591

*Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge large-scale replication of a randomized controlled trial. *American Educational Research Journal*, *52*(3), 516–546. https://doi.org/10.3102/0002831214565787

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)*. http://ies.ed.gov/ncee/edlabs

*Gorard, S., Siddiqui, N., & See, B. H. (2015). *Philosophy for children: Evaluation report and executive summary*. Education Endowment Foundation. https://files.eric.ed.gov/fulltext/ED581147.pdf

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

Hedges, L. 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*. WestEd. https://mpp.wested.org/wp-content/uploads/2013/05/mpp-ies-report.pdf

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. https://doi.org/10.1177/0735633117748416

*Hodgen, J., Adkins, M., Ainsworth, S., & Evans, S. (2019). *Catch Up® Numeracy: Evaluation report and executive summary*. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Catch_Up_Numeracy.pdf

*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. https://doi.org/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. https://doi.org/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*. 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*. 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.

Kennedy, A. (2014). Understanding continuing professional development. *Professional Development in Education*, *40*(5), 688–697. https://doi.org/10.1080/19415257.2014.955122

*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. https://doi.org/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*(Suppl. 1), 188–208. https://doi.org/10.1080/19345747.2015.1116031

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

*Kutaka, T. S., Smith, W. M., Albano, A. D., Edwards, C. P., Ren, L., Beattie, H. L., Lewis, J., Heaton, R. M., & 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. https://doi.org/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. https://doi.org/10.9734/BJESBS/2014/7259

*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, Germany*. Humboldt University. http://doc.renlearn.com/KMNet/R003476325GEFECD.pdf

*Lenard, M., & Rhea, A. (2019). *Adaptive Math and Student Achievement: Evidence from a randomized controlled trial of DreamBox Learning* [Paper presentation]. Annual Meeting of the Society for Research in Effective Education, Washington, DC, United States.

Lein, A., Jitendra, A., & Harwell, M. (2020). Effectiveness of mathematical word problem solving interventions for students with learning disabilities and/or mathematics difficulties: A meta-analysis. *Journal of Educational Psychology*, *112*(7), 1388–1408. https://doi.org/10.1037/edu0000453

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. https://doi.org/10.1007/s10648-010-9125-8

Lipsey, M. W. (2019). Identifying potentially interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 141–151). Russell Sage Foundation. https://doi.org/10.7758/9781610448864.11

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

*Lloyd, C., Edovald, T., Morris, S., Kiss, Z., Skipp, A., & Haywood, S. (2015). *Durham shared maths project. Evaluation report and Executive summary*. 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. https://doi.org/10.3102/0162373719849044

*Magnolia Consulting. (2012). *A final report for the evaluation of Pearson's Waterford Early Learning program: Year 2*. https://files.eric.ed.gov/fulltext/ED501575.pdf

*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*(October), 101782. https://doi.org/10.1016/j.cedpsych.2019.101782

Morrison, J. R., Ross, S. M., & Cheung, A. C. (2019). From the market to the classroom: How ed-tech products are procured by school districts interacting with vendors. *Educational Technology Research and Development*, *67*(2), 389–421. https://doi.org/10.1016/j.cedpsych.2019.101782

*Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G., & Barton, A. (2016). *ReflectED: Evaluation report and executive summary*. Education Endowment Foundation. https://files.eric.ed.gov/fulltext/ED581262.pdf

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (in press). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly*.

Nelson, G., & McMaster, K. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. *Journal of Educational Psychology*, *6*, 1001-1022. https://doi.org/10.1037/edu0000334

*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). U.S. Department of Education. https://doi.org/10.2139/ssrn.2511347

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

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

*Parker, D. C., Nelson, P. M., Zaslofsky, A. F., Kanive, R., Foegen, A., Kaiser, P., & Heisted, D. (2019). Evaluation of a math intervention program implemented with community support. *Journal*

of Research on Educational Effectiveness, *12*(3), 391–412. https://doi.org/10.1080/19345747.2019.1571653

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. (2021). *Data and program files associated with the publication: Effective Programs in Elementary Mathematics: A Meta-Analysis*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E130284V2

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programs. *American Educational Research Journal*, *48*(4), 996–1025. https://doi.org/10.3102/0002831211410864

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. https://doi.org/10.3102/0034654319877153

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*(April), 22–34. https://doi.org/10.1016/j.learninstruc.2018.01.009

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. *Journal of Educational Research*, *109*(5), 491–502. https://doi.org/10.1080/00220671.2014.992581

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

*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 presentation]. Annual Meeting of the Society for Research on Effective Education, Washington, DC, United States.

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

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

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

Rohrbeck, C.A., Ginsburg-Block, M.D., Fantuzzo, J.W., & Miller, T.R. (2003). Peer-assisted learning interventions wih elementary school students: A meta-analytic review. *Journal of Educational Psychology*, *94*(20), 240-257. https://doi.org/10.1037/0022-0663.95.2.240

*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. https://doi.org/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*. Education Endowment Foundation.

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

*Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J., Kibrick, M., Richland, L. E., Tran, N. A., Schneider, S. H., Duran, L., & 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. https://doi.org/10.1080/19345747.2013.856978

*Rutt, S., Easton, C., & Stacey, O. (2014). *Catch Up® Numeracy: Evaluation report and executive summary*. 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*(Novemebr), 158–172. https://doi.org/10.1016/j.edurev.2016.07.003

*Schoen, R. C., LaVenia, M., Tazaz, A., Farina, K., Dixon, J. K., & Secada, W. G. (2020). *Replicating the CGI experiment in diverse environments: Effects on grade 1 and 2 student mathematics achievement in the first program year (Research Report No. 2020–02)*. Florida State University. https://doi.org/10.33009/fsu.1601237075

*Schwartz, P. (2020). *Raising the bar district-wide using Symphony Math*. Hanover, NH: Symphony Learning.

*See, B. H., Morris, R., Gorard, S. G., & Siddiqui, N. (2018). *Maths Counts. Evaluation report and executive summary*. Education Endowment Foundation. 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), https://doi.org/10.1177/2332858419850482

Slavin, R. E. (2017). Instruction based on cooperative learning. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 388–404). Routledge.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*(3), 427–515. https://doi.org/10.3102/0034654308317473

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1-26. https://doi.org/10.1016/j.edurev.2010.07.002

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*(2), 839–911. https://doi.org/10.3102/0034654308330968

*Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating Math Recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement. *American Educational Research Journal*, *50*(2), 397–428. https://doi.org/10.3102/0002831212469045

*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 presentation]. Biennial Meeting of the Society for Research in Child Development, Montreal, Quebec, 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. https://doi.org/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*. Education Endowment Foundation.

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

*Styers, M., & Baird-Wilkerson, S. (2011). *A final report for the evaluation of Pearson's focusMATH Program*. 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. https://doi.org/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*. Education Endowment Foundation.

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*. Education Endowment Foundation.

*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*. 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. https://doi.org/10.1080/14794802.2013.797746

*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 presentation] Society for Research on Effective Education, Washington, DC, United States.

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*(6), 1251–1284. https://doi.org/10.3102/0002831212462736

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

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

*Wang, H., & Woodworth, K. (2011a). *A randomized controlled trial of two online mathematics curricula* [Paper presentation] Annual Meeting of the Society for Research on Effective Education, Washington, DC, United States.

*Wang, H., & Woodworth, K. (2011b). *Evaluation of Rocketship Education's use of DreamBox Learning's online mathematics program*. SRI International.

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. https://doi.org/10.1007/s10648-015-9321-7

Webb, N. M. (2008). Learning in small groups. In T.L. Good (Ed.), *21st Century education: A reference handbook* (pp. 203–211). Sage. https://doi.org/10.4135/9781412964012.n22

*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. https://doi.org/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*. Center for Educational Policy Research, Harvard University.

What Works Clearinghouse. (2013). *What works in math*. https://ies.ed.gov/ncee/wwc/Math/

What Works Clearinghouse. (2020). *Standards handbook* (Version 4.1). Author.

*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). U.S. Department of Education.

Wolf, R., Morrison, J.M., Inns, A., Slavin, R. E., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness*, *13*(2), 428–447. https://doi.org/10.1080/19345747.2020.1726537

*Wood, T., Mazzocco, M. M., Calhoon, M. B., Crowe, E. C., & Connor, C. M. (2020). The effect of peer-assisted mathematics learning opportunities in first grade classrooms: What works for whom? *Journal of Research on Educational Effectiveness*, *13*(4), 61–624. https://doi.org/10.1080/19345747.2020.1772422

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

*Wright, W., Dorsett, R., Anders, J., Buzzeo, J, Runge, J., & Sanders, M. (2019). *Improving working memory. Evaluation report and executive summary*. 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. https://doi.org/10.1080/02 796015.2007.12087933

Authors

MARTA PELLEGRINI is a fixed-term researcher at the University of Florence, Italy. Her research interests include evidence-based education and systematic reviews.

CYNTHIA LAKE is a senior research associate at the Center for Research and Reform in Education at Johns Hopkins University, Baltimore, Maryland. Her research interests include comprehensive school reform, school improvement, and educational research methods and statistics.

AMANDA NEITZEL is an assistant research scientist at the Center for Research and Reform in Education at Johns Hopkins University, Baltimore, Maryland. Her research interests include school-based health models, program evaluation, and systematic reviews.

ROBERT E. SLAVIN is the director of the Center for Research and Reform in Education at Johns Hopkins University, Baltimore, Maryland. He has authored or co-authored more than 300 articles and book chapters on such topics as cooperative learning, comprehensive school reform, research reviews, evidence-based reform, and vision and learning.