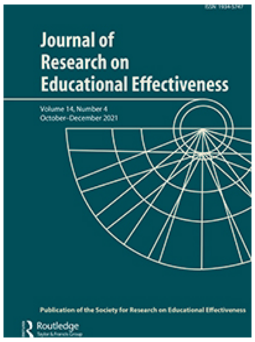


Journal of Research on Educational Effectiveness

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

# Heterogeneity in Mathematics Intervention Effects: Evidence from a Meta-Analysis of 191 Randomized Experiments

Ryan Williams, Martyna Citkowicz, David I. Miller, Jim Lindsay & Kirk Walters

**To cite this article:** Ryan Williams, Martyna Citkowicz, David I. Miller, Jim Lindsay & Kirk Walters (2022): Heterogeneity in Mathematics Intervention Effects: Evidence from a Meta-Analysis of 191 Randomized Experiments, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2021.2009072

**To link to this article:** https://doi.org/10.1080/19345747.2021.2009072

View supplementary material

Published online: 21 Jan 2022.

Submit your article to this journal

Article views: 1

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uree20

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2021.2009072

Routledge
Taylor & Francis Group

Check for updates

THEORY, CONTEXTS, AND MECHANISMS

# Heterogeneity in Mathematics Intervention Effects: Evidence from a Meta-Analysis of 191 Randomized Experiments

Ryan Williams[a], Martyna Citkowicz[a], David I. Miller[a], Jim Lindsay[a], and Kirk Walters[b]

[a]American Institutes for Research, Chicago, Illinois, USA; [b]WestEd, San Francisco, California, USA

**ABSTRACT**
Since the standards-based education movement began in the early 1990s, mathematics education reformers have developed and evaluated many interventions to support students in mastering more rigorous content. We conducted a systematic review and meta-analysis of U.S. PreK-12 mathematics intervention effects from 1991 to 2017 to study sources of heterogeneity. From more than 9,000 published and unpublished study reports, we found 191 randomized control trials that met our inclusion criteria, with 1,109 effect size estimates representing more than a quarter of a million students. The average effect size on student mathematics achievement was 0.31, with wide heterogeneity of most effects ranging from −0.60 to 1.23. Two modeling approaches—meta-regression and machine learning—provided converging evidence that outcome measure type (researcher-created vs. standardized) and technology delivery (vs. teacher or interventionist delivery) were predictors of effect size. Intervention type, intervention length, grade level, and publication year were also identified as potentially explanatory factors.

**ARTICLE HISTORY**
Received 11 September 2020
Revised 30 August 2021
Accepted 7 October 2021

**KEYWORDS**
Mathematics interventions; randomized experiments; meta-analysis

Improving the education of U.S. youth in the disciplines of science, technology, engineering, and mathematics (STEM) is a well-documented, widely endorsed federal policy priority (Schneider, 2021; White House, 2012). Underlying the advocacy for improved STEM education is the shared understanding that the numbers of STEM-related jobs are growing at much higher rates than jobs in non-STEM fields, and this trend is expected to continue (Fayer et al., 2017). Yet too few young Americans attain postsecondary degrees in STEM fields to meet this demand (Change the Equation, 2015; Langdon et al., 2011). Learning and applying STEM concepts can also be critically important to conducting tasks in non-STEM jobs as well as being capable, knowledgeable members of society (Zollman, 2012).

**CONTACT** Ryan Williams rwilliams@air.org American Institutes for Research, 10 S. Riverside Plaza, Suite 600, Chicago, IL 60606, USA.

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/19345747.2021.2009072

© 2022 Taylor & Francis Group, LLC

Mathematics is a foundation upon which STEM learning takes place. Traditional PreK-12 instruction in the United States typically teaches children basic math concepts (e.g., counting, measurement, basic arithmetic) before teaching concepts in the other STEM fields. Mathematics also provides the language and tools needed for understanding concepts and applications in other STEM fields (Basista & Mathews, 2002; Frykholm & Meyer, 2002). Improving children's understanding of mathematics is therefore an important goal for policymakers and educators who seek to ensure that American youth can later attain STEM-related jobs, the domestic supply of which outpaces the production of qualified graduates (President's Council of Advisors on Science and Technology, 2012). Policymakers also emphasize the importance of critical thinking and analyzing data to a well-rounded civics education (National Center for Education Statistics, 2020). Hence, mathematical proficiency is also important to developing an informed, productive citizenry.

Our study responds to these needs through a systematic review and meta-analysis of randomized controlled trials in U.S. PreK-12 mathematics education between 1991 and 2017. We focus on understanding the *heterogeneity* of intervention effects, identifying what types of mathematics intervention work, for whom, and under what conditions.

## Efforts to Improve Student Mathematics Achievement

Current U.S. student achievement in mathematics is lackluster at best. Despite some improvements between 1990 and 2009 on the National Assessment of Education Progress (NAEP), scores have plateaued since then and only 24% of students were at or above NAEP proficiency in mathematics by the time they graduated high school (National Center for Education Statistics, 2021). The United States ranked 32nd out of 41 industrialized nations for 15-year-olds' mathematics performance on the 2018 Program for International Student Assessment (Organisation for Economic Co-operation and Development, 2021). High school students who did not reach Algebra II were required to take remedial mathematics courses in college (Achieve, 2014). Having to take remedial mathematics courses in college is a significant barrier to enrolling and succeeding in STEM-related courses at the university level (Calcagno & Long, 2008). Thus, the shortages in the STEM workforce represent a more complex, systemic PreK–16 problem, of which mathematics plays a key role.

Public initiatives have emerged to address this problem, beginning as far back as the *A Nation At Risk* report in 1983 (National Commission on Excellence in Education, 1983). The National Council of Teachers of Mathematics (NCTM), for example, published two influential sets of standards in 1989 and 1991 that specified more rigorous content and process standards, curricula, and assessments for prekindergarten through Grade 12 education in mathematics (NCTM, 1989, 1991). NCTM (2000, 2007) and the National Research Council (Kilpatrick et al., 2001) continued the push for more rigorous content and process standards in the 2000s, followed more recently by the Common Core State Standards Initiative (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010).

Educators and policymakers seeking programs or interventions that improve student outcomes in mathematics education likely ask two fundamental questions: (1) Which programs are found to produce positive impacts on student outcomes? and (2) Under what conditions do those programs produce positive impacts? The first question is one of internal validity (i.e., Can one attribute changes in outcomes solely to the program?), whereas the second question is one of external validity (i.e., Do the findings generalize to other settings, outcomes, intervention features, and populations?). The field of education research has made significant progress in designing studies that optimally identify causal effects of programs and policies, exemplified by the Institute of Education Sciences' (IES's) emphasis on high-quality randomized controlled trials (RCTs) during the past two decades. Work remains on Question 2, identifying the conditions under which intervention effects vary in mathematics education. Without an understanding of intervention effect heterogeneity, we know little about conditions and contexts to which a study's findings generalize. The issue of generalizability is of utmost importance to the ultimate consumers of impact studies—practitioners. As much as they might be interested in whether an intervention showed impacts, they also want to know whether such findings are relevant to their students and schools. Moreover, the mathematics education literature provides abundant evidence for effect heterogeneity, as detailed next.

## Indications of Effect Heterogeneity for Mathematics Interventions

Heterogeneity in mathematics intervention effects is often found *within* primary studies. For example, in one cluster randomized trial of Kentucky Virtual Schools' hybrid program, researchers found little evidence of a program effect on the mathematics achievement of Grade 9 students (Cavalluzzo et al., 2012). However, the results of a sensitivity test revealed that average effects varied by study setting, with an average program impact estimate of −0.25 standard deviations in nonrural schools and an average impact estimate of 0.00 standard deviations in rural schools. Another example study evaluated the effects of the Tier 2 mathematics intervention on Grade 1 and Grade 2 students' scores on the Texas Early Mathematics Inventories (Bryant et al., 2008). The study found a statistically significant positive effect for Grade 2 students ($b = 0.19$, $p < .05$) but a small and nonsignificant estimate for Grade 1 students ($b = 0.04$, $p > .05$). The study also observed that Tier 2 students in Grade 2 benefited more than Tier 1 students in the same grade. Looking at variation in intervention effects by outcome subtests, the study found a significant positive effect only for the addition and subtraction combinations subtest ($b = 0.21$, $p < .05$).

The What Works Clearinghouse (WWC, 2020) also illustrates evidence for heterogeneity in mathematics intervention effects. The WWC focuses on reporting average impacts of specific interventions and educational practices for general student populations and specific subgroups. For instance, Cognitive Tutor® is a widely used and studied algebra intervention (Pane et al., 2014). For high school students, the WWC indicates that effects of this intervention are mixed, with an overall standardized mean difference effect size of −0.02 ($p > .05$) across the six intervention reports that met WWC standards with or without reservations. For middle school students, however, Cognitive Tutor® had positive results (effect size = 0.39, $p < .05$) for one study that met

WWC standards with or without reservations. WWC practice guides also show the levels of evidence supporting practices that might be embedded within interventions, such as for a recent practice guide on assisting elementary school students struggling with mathematics (e.g., Fuchs et al., 2021). The WWC's reporting of study findings allows for the manual inspection of how effects may vary across populations, but the systematic examination of intervention effect heterogeneity falls outside of the WWC's current scope.

Meta-analysis—the synthesis of quantitative findings across studies—provides tools to systematically examine effect heterogeneity. For example, a broad meta-analysis on teacher coaching programs across different content areas found that effects on teacher instruction and student achievement were smaller in larger-scale evaluation trials than smaller-scale trials (Kraft et al., 2018). The authors argued that these differences in average effects partly stem from the challenges in scaling up promising programs to new implementation contexts. The results led the authors to suggest ways in which program developers can approach these scale-up challenges, including training a diverse group of coaches and building teacher buy-in to implement a program. Meta-analyses on STEM education have also illustrated effect heterogeneity, including both in STEM broadly and in mathematics specifically. For instance, one meta-analysis investigated STEM-focused professional development programs and curricular materials (Lynch et al., 2019). Effects on student achievement were strongest when new curricular materials were combined with professional development, when programs focused on teachers' content knowledge and pedagogical content knowledge, and when teachers could meet to trouble-shoot implementation challenges. For mathematics education specifically, meta-analyses have found various moderators of intervention effects, such as the following examples:

- Treatment duration for mathematics intelligent tutoring systems (smaller effects for interventions lasting one year or longer; Steenbergen-Hu & Cooper, 2013).
- Student socioeconomic status (SES) for early numeracy interventions (smaller effects for low-SES students; Nelson & McMaster, 2019).
- Instructional group size for Tier 2 mathematics interventions for students with mathematics difficulties (larger effects for small groups of two or three students, relative to one-on-one instruction; Jitendra et al., 2021).

Several meta-analyses have also found larger effects for researcher-generated than standardized achievement measures (see Wolf, 2021 for an analysis of WWC data and a broader review across meta-analyses in education), including for meta-analyses of technology-enhanced mathematics interventions (Li & Ma, 2010), science education interventions (Taylor et al., 2018), and STEM-focused professional development programs (Lynch et al., 2019).

## Meta-Analytic Framework and Research Questions

Our study aimed to extend the systematic exploration of heterogeneity in mathematics intervention effects. Related synthesis efforts have typically focused on specific types of mathematics interventions (e.g., intelligent tutoring systems, professional development) or student populations (e.g., students with disabilities, elementary school students). These meta-analyses provide in-depth findings on specific aspects of PreK-12 mathematics education, but they provide less insight on how intervention effects across these areas contrast with each other. To address this limitation, we conducted a broad systematic review and meta-analysis of randomized experiments of interventions designed to improve mathematics learning among U.S. PreK-12 students, including studies published between 1991 and 2017.

We used Cronbach's (1982) UTOS (units, treatments, outcomes, and settings) model for generalizability as an organizing framework for our study. The UTOS framework refers to samples or units (Us), interventions or treatments (Ts), outcomes (Os), and settings (Ss) as features of studies. Primary study researchers are limited in the Us, Ts, Os, and Ss they can feasibly examine in any specific evaluation. However, researchers usually want to (implicitly or explicitly) generalize to a wider set of study characteristics that were not directly studied. In Cronbach's model, assessing generalizability involves determining how well a study's intersection of UTOS characteristics can be extrapolated to the "domain of application." Since Cronbach's initial publication in 1982, several meta-analysts have found the model a useful framework for calling attention to the key categories of moderators when examining effect heterogeneity (e.g., Ahn et al., 2012; Aloe & Becker, 2009; Becker, 2017).

Aloe and Becker (2009) extended the UTOS framework in a meta-analytic context to include variation due to methodological differences (Ms) in study designs, yielding the MUTOS framework (see also Becker, 2017). The M category is qualitatively different from the other categories because it usually represents nuisance variation (e.g., effect size computation details, variation in design attributes, approaches to measurement). For example, a randomized controlled trial and a quasi-experimental design could aim to estimate effects for the same intervention, student population, outcome, and setting, with the only difference being selection bias. Though selection bias is an important issue to mitigate, researchers usually do not otherwise care about it as a substantive feature of the intervention or its effect on students (other than potentially biasing estimation of the target effect).

The breadth of the MUTOS framework was well suited to our meta-analytic research questions:

1. How heterogeneous are mathematics intervention effects for U.S. PreK-12 students?
2. What factors contribute to this mathematics intervention effect heterogeneity?

## Method

Our study followed Becker's (2017) six recommended steps for using the MUTOS framework to investigate effect heterogeneity in meta-analyses:

1. Identify the desired target of inference by defining the relevant MUTOS characteristics
2. Code study features using MUTOS
3. Descriptively evaluate the diversity for each component of MUTOS

4. Assess overall heterogeneity of effects
5. Evaluate empirical variation in effects for each component of MUTOS
6. Assess connections, or generalizability, to desired domain of application

The first three steps correspond to our systematic review to identify eligible studies, code their study characteristics, and conduct a descriptive analysis of coding frequencies. The last three steps correspond to our meta-analysis to use statistical methods to analyze variation in standardized mean differences, identify specific sources of heterogeneity, and interpret the results based on key categories relevant to considering study generalizability.

## Systematic Review

Our study started with a systematic review to search for, screen, and code mathematics intervention studies. Our literature search and retrieval process for our systematic review of mathematics intervention studies is presented in Figure 1, and our inclusion criteria is defined in Table 1, corresponding to Step 1 from Becker (2017).

### Literature Search

We first conducted electronic database searches of ERIC, Education Source, PsycINFO, Psychology & Behavioral Sciences Collections, SocINDEX, Academic Search Premiers, JSTOR, WorldCat, and the NBER Working Papers. The search was limited to English language–only studies published between January 1991 and August 2017, focusing on mathematics-related topics in Grades PreK–12 (see the supplemental materials for a complete list of search terms). We also conducted a gray, or unpublished, literature search by searching the U.S. Department of Education websites, such as the WWC, and websites of research organizations, such as Mathematica and the National Research and Development Center on Cognition and Mathematics Instruction (see the supplemental materials for a complete list). After removing duplicates, the database and gray literature searches yielded 9,384 titles.

### Screening

We conducted study screening in three stages: (1) title and abstract screening, (2) full-text screening, and (3) methods screening. Title and abstract screening focused on determining whether the appropriate interventions, outcomes, and samples were included in the study (Criteria 1, 3, and 4 in Table 1). Full text screening focused on confirming that the appropriate interventions, outcomes, and samples were included in the study as well as determining whether an eligible control group was used, whether the study was written in English, and whether the study took place in the United States or its territories (Criteria 1, 2, 3, 4, and 6 in Table 1). Methods screening focused on determining whether the study used uncompromised random assignment, was free of $N = 1$ confounds,[1] and whether enough information was provided to calculate an effect size

[1] $N = 1$ confounds occur when the intervention or comparison group contains only one study unit.

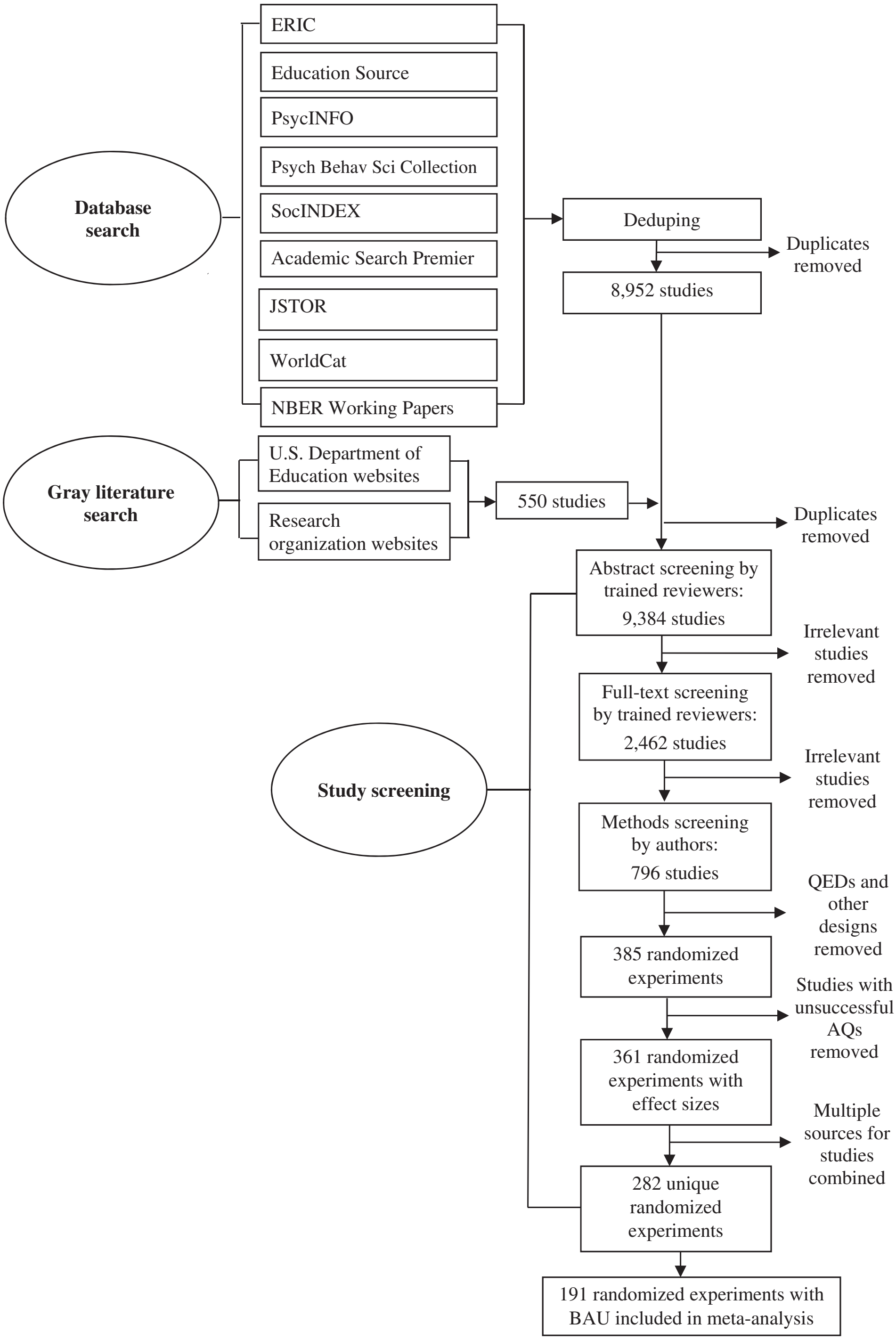


**Figure 1.** Mathematics interventions systematic review literature search and retrieval process.

**Table 1.** Eligibility criteria.

| Criteria | |
|---|---|
| 1. | Included at least one mathematics intervention, defined as an intervention, strategy, or program designed specifically to improve the teaching or learning of mathematics. Professional development, curricula, after school programs, games, and spatial reasoning strategies were all included so long as the goal of the intervention was to improve mathematics teaching or learning. |
| 2. | Conducted a randomized controlled group trial without $N = 1$ confounds. |
| 3. | Included a sample of students in Grades PreK–12 in the United States or its territories. |
| 4. | Evaluated at least one measure of mathematics learning or knowledge (including measures of acquisition, maintenance, or achievement). |
| 5. | Study provided sufficient information to calculate an effect size and its variance. |
| 6. | Written in English. |
| 7. | Published in 1991 or later. |
| 8. | Included a business-as-usual (BAU) control group. |

*Note.* $N = 1$ confounds occur when the intervention or comparison group contains only one study unit. Criterion 8 was applied at the analysis stage, after all coding was completed.

estimate and variance (Criteria 2 and 5 in Table 1). Rather than screening on study attrition or baseline equivalence, our approach treated those characteristics as potential methodological moderators, which are presented in the results below.

Three trained reviewers screened the titles and abstracts of all 9,384 studies. Five trained reviewers screened the full text of the 2,462 studies that made it to the second stage. The authors screened for the study methods of the 796 studies that made it to the third stage. This overall screening process yielded 191 unique randomized controlled trials that met all eligibility criteria, had sufficient information to extract effect sizes, and had at least one business-as-usual comparison group.[2] Training reviewers and authors included assigning the same set of 10 studies to all reviewers and meeting to discuss any discrepancies to align on the inclusion criteria. This training took place twice per stage prior to assigning studies for screening. The second author also met with reviewers weekly at all stages of screening to prevent screening drift over time.

Studies were dual screened at each stage using a random sampling strategy to continuously monitor screening and coding deviations from the protocol. Ten percent of studies were dual screened at the abstract and methods stages, and 30 percent of studies were dual screened at the full-text stage. Any discrepancies were resolved by one of the authors. Interrater reliability was 0.77 at the abstract stage, 0.88 at the full-text stage, and 0.78 at the methods stage. Most discrepancies resulted in an adjudication of exclusion by the authors as reviewers were instructed to err on the side of inclusion.

All screening took place in a Microsoft Access® database created for this project. Reviewers answered questions using the criteria defined in Table 1. A "No" response from reviewers to any of the questions excluded the study from further review. If the reviewers responded "Yes" or "Do Not Know" to all questions at a given stage, the study moved to the next stage. Studies that made it through all three stages of screening made it to the coding phase.

[2]Our broader project coded 282 RCT studies, but this manuscript's analyses focus on the 191 studies with at least one business-as-usual (BAU) control group. We excluded 91 studies that had only alternative treatment comparisons (and no BAU group) due to the complexity and lack of clear methodological guidance on analyzing such studies (e.g., the effect size could be positive or negative if the choice of the "main" intervention is not clear).

## Coding

As noted earlier, we used the MUTOS framework to guide our approach for coding study characteristics as potential moderators. We selected codes for each MUTOS category by consulting the empirical literature on what study characteristics have yielded heterogeneous intervention effects, including in individual primary studies or meta-analyses on student achievement in mathematics or other STEM fields. We also consulted with the study's mathematics content experts to ensure the selected codes were theoretically and empirically relevant to our research questions, separately for each MUTOS category. Additionally, we considered the reporting prevalence of specific study features, aiming to balance coding comprehensiveness, feasibility, and utility for final analysis.

Six coders coded the 191 eligible studies, with 10% of the studies being dual coded. Any coding discrepancies were discussed by the two coders, and, if not resolved, the second author intervened and made a final decision. The second author met with the coders weekly to discuss any questions and prevent coding drift over time. We coded study-level information related to publication and study design; sample characteristics, such as sample size and demographics; intervention characteristics related to broad type (i.e., curriculum, pedagogical/instructional, or supplemental time), training, and delivery; outcome measure information, such as type and domain; setting information, such as urbanicity and locale; and summary statistics to calculate effect sizes. A copy of the codebook is included in the online supplemental materials. Table 2 also lists these codes along with descriptive frequencies.

For the intervention codes developed for this study, we focused on three broad intervention types: curriculum, pedagogical/instructional, and supplemental time. Curriculum interventions were those where the primary component was an experimental curriculum of some sort, which could have included traditional classroom, online, blended, or reform oriented curricular materials. Pedagogical and instructional interventions were those that focused primarily on improving mathematics instruction, such as professional development, coaching, and other teacher training interventions. Supplemental time interventions were those that programs that were intended as "add-ons" to standard curriculum and instructional activities, such as tutoring, double-periods, afterschool supports. In cases, where it was unclear which intervention type was the primary intervention type, we consulted with a mathematics content expert to help make the determination.[3]

The Access database used for coding had a hierarchical structure such that study-level information was coded first; followed by the intervention name and corresponding characteristics; then sample and setting information; outcome information; and finally effect size information. We coded all eligible information, including all interventions, samples, outcomes, and effects. For example, if a study examined one intervention for two separate samples for whom they measured three different mathematics outcomes at two time points (directly after the intervention and follow-up), we coded one intervention page, two sample pages, three outcome pages per sample, and two effect size pages per outcome, resulting in a total of 12 effect sizes. If studies reported on both an overall

[3]We recognize that blended or hybrid intervention types may be important and missing parts of this broad typology and we reflect on this issue in the limitations section.

**Table 2.** Characteristics of 191 included experiments (1,109 effect sizes).

| Characteristic | *m* | *k* | Mean (SD) | Missing (%) |
|---|---|---|---|---|
| Methods characteristics | | | | |
| Random assignment level | | | | |
| Student | 93 | 547 | 49% | 0% |
| Teacher/Classroom | 67 | 379 | 34% | 0% |
| School | 33 | 183 | 17% | 0% |
| District | 0 | 0 | 0% | 0% |
| Published journal article | 117 | 764 | 69% | 0% |
| Attrition | | | | |
| Low[a] | 45 | 255 | 23% | 0% |
| High | 10 | 31 | 3% | 0% |
| Insufficiently reported to assess | 146 | 823 | 74% | 0% |
| Sample characteristics | | | | |
| Grade level | – | – | 3.32 (2.93) | 4% |
| Prekindergarten | 18 | 82 | 8% | 4% |
| Elementary school | 112 | 767 | 72% | 4% |
| Middle school | 63 | 239 | 23% | 4% |
| High school | 28 | 85 | 8% | 4% |
| Demographics | | | | |
| % Male | – | – | 52% (14%) | 30% |
| % Special education | – | – | 20% (28%) | 72% |
| % English language learner | – | – | 22% (24%) | 65% |
| % Economically disadvantaged | – | – | 57% (24%) | 58% |
| % White | – | – | 40% (27%) | 41% |
| % Hispanic | – | – | 25% (23%) | 43% |
| % Black | – | – | 32% (23%) | 40% |
| % Asian | – | – | 6% (10%) | 59% |
| Intervention characteristics | | | | |
| Intervention type | | | | |
| Curriculum | 83 | 443 | 40% | 0% |
| Pedagogical/Instructional | 85 | 553 | 50% | 0% |
| Supplemental time | 24 | 113 | 10% | 0% |
| Intervention content domain | | | | |
| Number sense and arithmetic | 90 | 642 | 67% | 13% |
| Rational numbers and fractions | 39 | 196 | 20% | 13% |
| Algebra and prealgebra | 57 | 269 | 28% | 13% |
| Geometry | 42 | 230 | 24% | 13% |
| Measurement, data, and statistics | 39 | 214 | 22% | 13% |
| Calculus and precalculus | 1 | 1 | 0% | 13% |
| Implementation fidelity | | | | |
| High | 41 | 395 | 72% | 50% |
| Medium | 22 | 114 | 21% | 50% |
| Low | 9 | 42 | 8% | 50% |
| Implementation training | | | | |
| None or not reported | 67 | 376 | 34% | 0% |
| One-time training | 58 | 353 | 32% | 0% |
| Infrequent ongoing training | 38 | 193 | 17% | 0% |
| Frequent ongoing training | 30 | 187 | 17% | 0% |
| Intervention delivery | | | | |
| Teacher | 110 | 608 | 55% | 0% |
| Technology | 65 | 375 | 34% | 0% |
| Interventionist | 52 | 380 | 34% | 0% |
| Number of hours | – | – | 23.55 (29.92) | 30% |
| <=1 h | 12 | 63 | 8% | 30% |
| >1 h and <= 4 h | 21 | 89 | 11% | 30% |
| >4 h and <= 20 h | 47 | 390 | 50% | 30% |
| >20 h | 45 | 235 | 30% | 30% |
| Number of weeks | – | – | 19.77 (17.91) | 6% |
| <=1 week | 11 | 61 | 6% | 6% |
| >1 week and <=4 weeks | 35 | 180 | 17% | 6% |
| >4 weeks and <=18 weeks | 59 | 392 | 38% | 6% |
| >18 weeks | 71 | 409 | 39% | 6% |

(*continued*)

**Table 2.** Continued.

| Characteristic | $m$ | $k$ | Mean (SD) | Missing (%) |
|---|---|---|---|---|
| Outcome characteristics | | | | |
| Outcome measure content domain | | | | |
| Number sense and arithmetic | 92 | 578 | 64% | 18% |
| Rational numbers and fractions | 40 | 190 | 21% | 18% |
| Algebra and prealgebra | 61 | 216 | 24% | 18% |
| Geometry | 47 | 164 | 18% | 18% |
| Measurement, data, and statistics | 45 | 157 | 17% | 18% |
| Calculus and precalculus | 0 | 0 | 0% | 18% |
| Outcome type | | | | |
| Standardized achievement measure | 107 | 470 | 42% | 0% |
| Researcher-generated measure | 122 | 637 | 57% | 0% |
| Course credits/Enrollment[b] | 2 | 2 | 0% | 0% |
| Outcome timing | | | | |
| Midstream during intervention | 14 | 33 | 3% | 0% |
| Immediate posttest | 172 | 898 | 81% | 0% |
| Follow-up posttest | 42 | 169 | 15% | 0% |
| Combination of time periods | 4 | 9 | 1% | 0% |
| Outcome-intervention alignment[c] | – | – | 0.89 (0.27) | 24% |
| Setting characteristics | | | | |
| Urbanicity | | | | |
| Suburban | 51 | 299 | 45% | 40% |
| Urban | 82 | 482 | 72% | 40% |
| Rural | 39 | 222 | 33% | 40% |
| U.S. geographic region | | | | |
| West | 35 | 206 | 22% | 16% |
| Midwest | 30 | 188 | 20% | 16% |
| Southwest | 41 | 260 | 28% | 16% |
| Northeast | 51 | 364 | 39% | 16% |
| Southeast | 64 | 326 | 35% | 16% |
| Publication year | – | – | 2010.68 (5.08) | 0% |
| 1990s | 12 | 50 | 5% | 0% |
| 2000s | 52 | 290 | 26% | 0% |
| 2010s | 127 | 769 | 69% | 0% |

*Note*. Percentages are based on the frequencies of effect sizes (e.g., 49% of effect sizes were from student-level RCTs, corresponding to $k = 547$ effect sizes across $m = 93$ studies). Percentages may sum to more than 100% for characteristics that are not mutually exclusive (e.g., a study could be conducted in both rural and urban settings and across multiple grade levels).
$m$ = number of studies, $k$ = number of effect sizes, Mean (SD) = average percentage for nonmissing values (and standard deviation for continuous moderators), Missing (%) = percentage of effect sizes that have missing values for that characteristic (e.g., such as studies reporting a "mathematics achievement" measure without specifying the outcome domain).
[a]The determination of low attrition was based on meeting the optimistic boundary for both low student-level and randomization-level attrition under What Works Clearinghouse Group Design Standards, Version 4.1 (What Works Clearinghouse, 2020).
[b]The course credits/enrollment category was combined with researcher-generated measures at the analysis stage.
[c]Outcome-intervention alignment was the proportion of overlap between the outcome domains covered in the outcome measure and the content covered in the intervention. For instance, the score would be 0.50 if the outcome measure covered number sense and basic operations, but the intervention focused only on number sense.

mathematics score and subscores, we prioritized including the subscores. However, we did not code individual items (e.g., when a mathematics score was made up of a single mathematics problem).

We also used all available sources to code each study. For example, if a study included a peer-reviewed article, report, and conference abstract, we used all three sources to code as much information as possible. If discrepancies arose across the sources, we discussed these discrepancies and generally trusted the peer-reviewed article or sources with the latest publication date to be the most accurate. We attempted to find all relevant sources the included studies. For example, for all WWC intervention reports, conference abstracts, executive summaries, and errata, we searched for the corresponding journal articles and reports first within our list of studies and, if we could not find a copy there, we searched the internet. All sources were linked via partial title names and then double checked manually to ensure proper linking of study sources.

## *Meta-Analysis*

### *Computing Effect Sizes*

We computed effect sizes to provide a common metric for synthesis across studies that measure outcomes on different scales. Effect sizes encode both the direction and the magnitude of the relationship between intervention and outcomes (Hedges & Olkin, 1985; Lipsey & Wilson, 2001). Specifically, we computed the standardized mean difference (SMD) effect size for all mathematics-related outcomes reported in each study. We used reported summary statistics, including means and standard deviations, $t$ tests, $F$ tests, $\chi^2$ tests, regression coefficients, and effect sizes in other metrics to compute the SMDs.[4] The equations for calculating the SMD, or converting other effect size metrics to the SMD, can be found in Borenstein, Hedges, Higgins, and Rothstein (2009).

We applied two adjustments to the SMDs and their variances. First, we used Hedges's (1981) small sample bias correction to the effect size estimate to the account for small studies. Second, we adjusted the effect size variances for clustering when the level of random assignment was at the cluster level (e.g., teachers or schools were randomly assigned to conditions), using formulas provided by Hedges (2007, 2011).

### *Meta-Analytic Models*

Our focal analyses used mixed-effects meta-regression models to investigate sources of effect heterogeneity. These models assumed that observed variation in effect sizes was due to fixed effects of moderators (e.g., intervention type), random effects of residual effect heterogeneity, and within-study sampling variance (Borenstein et al., 2009). Models were estimated using restricted maximum likelihood with the *metafor* package in the statistical software R (Viechtbauer, 2010).

To account for effect size dependencies (i.e., multiple effects per study), we used robust variance estimation to adjust the standard errors and degrees of freedom for regression coefficients, using the small-sample correction based on the Satterthwaite approximation (Tipton, 2015; Tipton & Pustejovsky, 2015) and the *clubSandwich* R package (Pustejovsky, 2018). The model specification using the *rma.mv()* function in the *metafor* package accounted for effect size dependencies based on both hierarchical or multilevel structures (i.e., subsamples nested within studies) and correlated, multivariate structures (i.e., multiple measures for the same sample). Pustejovsky and Tipton (2021) describe this approach in more detail (see our script 01_analysis.R for how we implemented this approach; see link below). We assumed a correlation of $r = .50$ for effect

[4]Appropriate summary statistics were not always available to calculate SMDs for all effects. We queried authors for the missing information, which yielded some success in obtaining the necessary data to calculate SMDs. Our response rate for queries was 42%.

sizes for multiple outcomes nested within the same sample; sensitivity analyses that varied this assumed correlation parameter showed robustness of the overall mean effect size, its standard error, and overall heterogeneity estimate, which we operationalized as the sum of the within- and between-study variance components. As expected, results showed some sensitivity to the relative partitioning of within-study and between-study heterogeneity, but this issue did not affect our central analyses which focused on overall heterogeneity (for further detail, see pages 9–10 and Table S1 in the supplemental materials).

We used multiple imputations to account for missing moderator data (e.g., missing racial demographics), as recommended by Pigott (2001; 2012). For imputing missing data, we used the *jomo* R package to account for the multilevel structure of the data (i.e., effects nested within studies; Quartagno et al., 2019) and aggregated results across 80 imputations, accounting for both the within- and between-imputation variance (Barnard & Rubin, 1999; Pustejovsky, 2017).

The data and code for these analyses available at https://osf.io/f9gud/?view_only=c97ba1316ff44606b8954d686e4d2d8b.

### Interpreting Meta-Analytic Results

We quantified heterogeneity using the model-based variance estimates (summing both the within- and between-study components) and 95% prediction intervals (i.e., estimated dispersion of the middle 95% of true underlying effects; Borenstein et al., 2017).[5] We also calculated the estimated percentage of true effects that were positive (greater than 0) or larger than practically important thresholds such as 0.10 or 0.25 standard deviations, as recommended by IntHout et al. (2016) and Mathur and VanderWeele (2019). We used cluster bootstrapped confidence intervals to quantify uncertainty in the heterogeneity estimates and the percentages of effect sizes above certain thresholds. The 10,000 bootstrap iterations sampled at the study level, not the effect size level, to account for effect size dependencies. We used the *boot* R package to generate bias-corrected and accelerated confidence intervals (Canty & Ripley, 2020; our analysis script 02_results.R available on the OSF website details our implementation).

We also computed post-estimation conditional means for categorical moderators (i.e., covariate-adjusted means for each intervention type, keeping other moderator values constant). For example, the conditional mean for curriculum interventions represents the model-predicted mean if the entire sample of effect sizes were about curriculum interventions while holding the other moderators in the meta-regression model constant (e.g., level of assignment, standardized vs. researcher-generated measure, intervention length). These predicted values therefore enable comparison of means while adjusting for potential confounds. The supplemental materials explain the computation of these conditional means in further detail.

[5]The prediction intervals estimated were based on a standard normal distribution: $PI = g \pm \tau(1.96)$, where $g$ is the estimated average effect and $\tau$ is the estimated between-effect standard deviation.

### Model Building Process

In all moderator analyses, we adjusted for potential methodological confounders such as level of random assignment; attrition; effect size computation details; publication status (published vs. unpublished); and outcome type (standardized vs. researcher-generated measure). Although they were not of central theoretical interest, these methods moderators could act as confounders, potentially biasing other moderator results of interest. Hence, we included them in all mixed-effects models, regardless of their statistical significance, as recommended by Tipton, Pustejovsky, and Ahmadi (2019). We had considered including outcome type (standardized vs. researcher-generated measure) as the part of the outcomes ("O") category in the MUTOS framework. However, we decided it is better positioned as part of the methods ("M") category because (a) not controlling for it could confound other moderators of interest (Wolf, 2021) and (b) it primarily reflects a methodological difference as opposed to a substantive difference like algebra versus geometry outcomes.

We first ran a mixed-effects meta-regression model with only methods ("M") moderators. We then ran four separate models, which each included a group of moderators for each UTOS component in addition to the methods moderators. For example, the model corresponding to the "O" component of UTOS (i.e., outcomes) included outcome moderators such as dummy codes for the outcome domain (e.g., algebra assessment) and timing (e.g., immediate or delayed posttest), in addition to the methods moderators. From these four models, we selected moderators with $p$-values less than .10 for inclusion into a combined model.

### Robustness of the MUTOS Model

As an exploratory sensitivity analysis, we also used a machine learning method called *random forests* as an algorithmic approach to model building (Breiman, 2001). The random forest algorithm is a powerful and flexible tool that can outperform simple linear regression in predicting outcomes, especially when moderators and effect sizes have complex relationships (e.g., nonlinearities and interactions). The MetaForest R package adapted this algorithm to the meta-analytic context of investigating which moderators best predict heterogeneity in effect sizes (van Lissa, 2017). The supplemental materials describe our application of this approach in more detail, including how we "fine-tuned" the random forest model based on van Lissa's (2020) recommendations. The random forest approach builds on simpler decision tree models such as Meta-CART (Li et al., 2020) while addressing several of their limitations such as their instability to slight data variations and tendency to overfit, yielding improved predictions[6] (van Lissa, 2017).

We used this machine learning approach to answer two main questions:

1. How does our linear meta-regression model (guided by the MUTOS theoretical framework) contrast with a machine learning model (guided by automated algorithms) in predicting effect sizes?

[6]We also conducted exploratory analyses using the Meta-CART package (Li et al., 2020), but the results indicated worse predictive performance compared to even standard linear meta-regression models.

2. To what extent do these two modeling approaches yield similar conclusions about the most important moderators?

### Selective Reporting Bias Analyses

The supplemental materials also detail our analytic approaches to diagnose and adjust for selective reporting bias (e.g., such as publishing only studies with statistically significant, favorable intervention effects). In short, we used three approaches: (a) comparison of unpublished versus published studies, (b) meta-regression to assess small-study effects, and (c) selection modeling. Despite the advances in selective reporting analytic methods, these approaches should be viewed as sensitivity analyses, rather than definitive, bias-corrected, estimates (Carter et al., 2019).

## Results

### Study Search Results

Using the literature search and retrieval process shown in Figure 1, we found 191 unique RCT studies that had at least one business-as-usual control group. These studies included more than a quarter million student participants. We extracted 1,109 effects from the 191 included studies, with a minimum of 1 effect size per study and a maximum of 48 (median = 4, mean = 5.76). The multiplicity of effect sizes came from studies having both multiple samples (median = 2, mean = 2.66) and multiple outcome measures within a study (median = 2, mean = 2.26).

### Descriptive Statistics About Study Characteristics

Table 2 provides summary information about the studies and their coded characteristics, organized around the MUTOS framework. Regarding methods characteristics, about half (49%) of effect sizes came from studies with individual-level random assignment, whereas the other half came from studies with cluster-level assignment of teachers/classrooms (34%) or schools (17%). Notably, information about attrition was usually not reported (74% of the time) in enough detail to assess both student-level and assignment-level attrition rates against the WWC's (2020) attrition standards.

The included study samples were demographically diverse: 40% of students were White, 32% were Black, 25% were Hispanic, 6% were Asian, and 57% were economically disadvantaged (as generally indicated by free or reduced-price lunch status) when those demographic statistics were reported; percentages were weighted by the number of effect sizes. However, this demographic information was often not reported (missing data rates varied from 30% to 72%). The earliest grades among the PreK–12 grade band were overrepresented. For example, 72% of samples included elementary school students compared with 8% for high school students. Hence, for these RCTs, the U.S. PreK–12 mathematics education research community has largely prioritized early childhood and elementary school learning.

The mathematics interventions were most often instructional or pedagogical strategies (50%) and replacement curriculum units (40%) and least often supplemental time interventions such as tutoring outside of normal classroom instructions (10%).[7] Information about implementation fidelity was often not reported (50%), but when it was reported, the study authors usually judged it to be high (72%). Most interventions lasted longer than 4 h (80%) and longer than 4 weeks (77%).

Regarding outcome characteristics, most measures were administered immediately following completion of the intervention (81%), and most measures were researcher-generated (57%) rather than standardized measures (42%). The outcome content domain also demonstrated the field's focus on young children—most measures assessed understanding of number sense and basic arithmetic (64%). The next most common category was algebra or prealgebra measures (24%).

The studies were distributed geographically across all major U.S. regions (e.g., West, Northeast), usually within urban settings (72% of the time when information on the locale was reported), although suburban and rural settings were also common (45% and 33%, respectively[8]). Most studies were published between 2010 and 2019 (69%).

## *Meta-Analytic Results*

The unadjusted random effects average effect was 0.31 ($SE = 0.03$, $df = 170.36$, $p < .01$, 95% CI [0.26, 0.37]), and heterogeneity was large ($\tau = 0.47$, 95% CI [0.37, 0.58], based on combining the between-study and within-study heterogeneity parameter estimates). The estimated middle 95% of true underlying effects (i.e., the 95% prediction interval) was between −0.60 to 1.23. Based on the average effect and overall heterogeneity, the probability that a random mathematics intervention effect has a positive impact is 75% (95% CI [71%, 78%]).[9] The probability for having an effect of at least 0.10 and 0.25 standard deviations is 68% (95% CI [64%, 71%]) and 55% (95% CI [52%, 59%]), respectively.

As shown in Figure 2 and Table 3, the unconditional means and distributions were similar across the three broad intervention types of curriculum ($g = 0.31$), pedagogical/instructional ($g = 0.32$), and supplemental time ($g = 0.35$) interventions. Importantly, however, other study characteristics could vary across intervention types (e.g., use of standardized versus research-generated measures), which could distort the interpretation of these unconditional means; the following results suggest that supplemental time interventions may have larger effects than the other two types after controlling for potential confounds.

As a first step in our exploration of moderators, we examined the effects of each of the MUTOS components, always controlling for the methods block (see Table 4 for the estimated residual heterogeneity values from the different moderator models). The methods block included outcome type (i.e., standardized versus researcher-generated measure); publication status; National Center for Education Evaluation and Regional

[7]The *primary* intervention type was always coded; thus, if a curriculum intervention also included some pedagogical strategies, it was only coded as a curriculum intervention.

[8]Studies often included schools from more than one than locale setting (e.g., urban and suburban), thus, these percentages sum to greater than 100%.

[9]This estimate assumed that the effect distribution is normally distributed with a mean of 0.31 and a standard deviation of 0.47 (see Mathur & VanderWeele, 2019). We used cluster bootstrapping sampling at the study level, not effect size level, to compute the confidence intervals (see the Methods section for further detail).

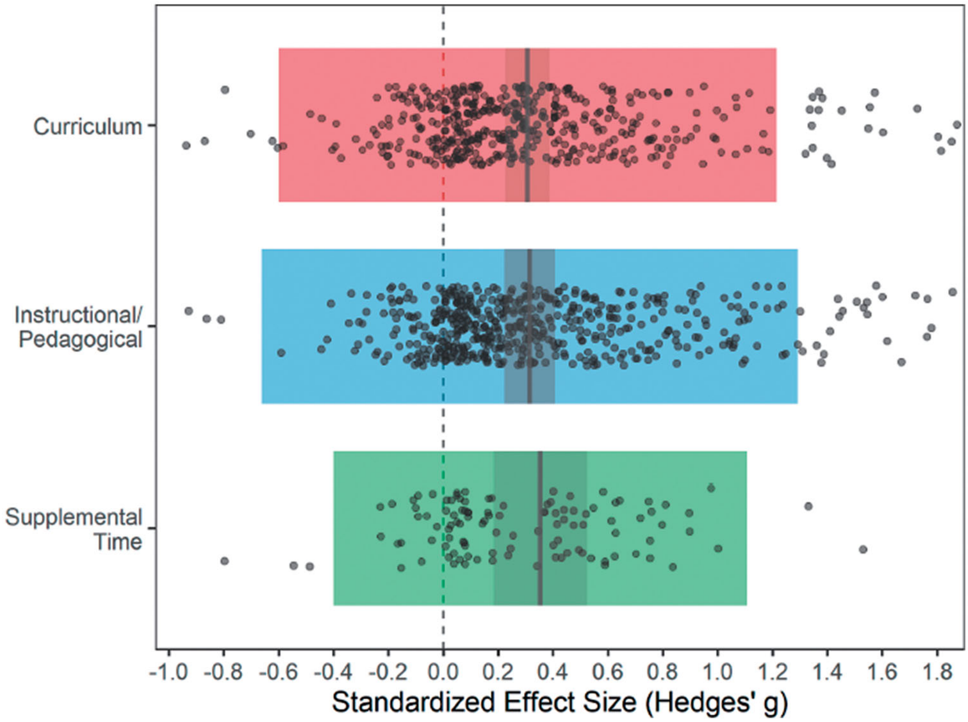


**Figure 2.** The effect size distribution by mathematics intervention type. The outer shaded boxes show the 95% prediction intervals, the estimated middle 95% of true underlying effects. The inner shaded boxes show the 95% confidence intervals, represented uncertainty in the overall mean estimate. The thick gray lines show the weighted means from random-effects models. This graph is intended as a descriptive summary of unconditional means (see the Results section text for more description of differences in conditional vs. unconditional means).

**Table 3.** Random-effects meta-analyses conducted separately by intervention type.

| Intervention Type | *g* | *SE* | *m* | *k* | *df* | *p* | $\tau$ | 95% Prediction interval |
|---|---|---|---|---|---|---|---|---|
| Curriculum | 0.31 | 0.04 | 83 | 443 | 70.93 | <.01 | 0.46 | [−0.60, 1.21] |
| Pedagogical/Instructional | 0.32 | 0.05 | 85 | 553 | 75.22 | <.01 | 0.50 | [−0.66, 1.29] |
| Supplemental Time | 0.35 | 0.08 | 24 | 113 | 19.76 | <.01 | 0.38 | [−0.40, 1.11] |

*Note*. These statistics come from random-effects meta-analyses estimated separately by mathematics intervention type. The standard errors were adjusted for effect size dependencies using robust variance estimation. $g$ = average effect size, $SE$ = standard error of the average effect sizes, $m$ = number of studies, $k$ = number of effect sizes, $df$ = degrees of freedom, $p$ = significance level for the mean being different from 0, $\tau$ = estimated standard deviation of the true underlying effect sizes, 95% prediction interval = estimated middle 95% of the true underlying effect sizes.

Assistance (NCEE) trail status; assumed correlation[10]; WWC attrition and baseline equivalence; and level of random assignment. Only one of these methods moderators was statistically significant: outcome type (see Table 5). Average effects were larger for research-generated than standardized measures ($g = 0.45$ and 0.15, respectively). All

[10]Assumed correlation reflects whether a correlation was imputed in calculating the effect size, which applies to three scenarios: (1) standard deviations for pre-post gain scores were reported, which had to be corrected; (2) the effect size was based on an ANCOVA *F*-test statistic, but the model $R^2$ value was not reported; and (3) the effect size was based on an unstandardized regression coefficient and its standard error, but the posttest standard deviation was not reported. In total, this designation applied to 10 of 1,109 effect sizes (1%).

**Table 4.** Meta-regression results for overall blocks of moderators.

| Model | Total $\tau$ | $\tau_B$ | $\tau_W$ | $p$ | $R^2$ |
|---|---|---|---|---|---|
| No moderators | 0.47 | 0.30 | 0.36 | – | – |
| Methods only | 0.46 | 0.30 | 0.35 | <.01[a] | 5.1% |
| Methods + Sample | 0.45 | 0.29 | 0.34 | .80 | 10.8% |
| Methods + Intervention | 0.45 | 0.29 | 0.35 | .21 | 7.5% |
| Methods + Outcome | 0.45 | 0.29 | 0.35 | .81 | 5.8% |
| Methods + Setting | 0.46 | 0.30 | 0.34 | .76 | 4.4% |
| Methods + Selected Moderators | 0.44 | 0.28 | 0.35 | .01 | 10.2% |

*Note*. The first column ($\tau$) is an estimate of the total residual effect heterogeneity (i.e., variability in the true underlying effect sizes not accounted for by included moderators). The total $\tau$ incorporates both the between-study heterogeneity $\tau_B$ and within-study heterogeneity $\tau_W$ where $= \sqrt{\tau_B^2 + \tau_W^2}$ . The relative partitioning of variance to $\tau_B$ versus $\tau_W$ should be interpreted cautiously because it depended heavily on the assumed within-sample correlation, though the total $\tau$ estimates were more robust (see Table S1 in the supplemental materials).
The $p$ values are based on multivariate Wald tests assessing whether the added group of moderators were significant. The $R^2$ is the percentage reduction in estimated effect variance compared with the random-effects model (first row). The reported heterogeneity values ($\tau$) represent standard deviations, rather than variances, so that they are on the same scale as effect size estimates. However, the $R^2$ values were based on reduction in variances ($\tau^2$).
[a]The $p$ value for the methods only moderators assessed the significance of methods moderators, not controlling for any other moderators. However, all other $p$ values assessed the significance of other groups of moderators (e.g., sample moderators) after controlling for methods moderators.

**Table 5.** Meta-regression results for methods-only moderators.

| Moderator | Mean | *SE* | *m* | *k* | *df* | *p* |
|---|---|---|---|---|---|---|
| Outcome type | | | | | 40.21 | <0.01 |
| Researcher-generated measure | 0.45 | 0.05 | 123 | 639 | 109.70 | |
| Standardized achievement measure | 0.15 | 0.05 | 107 | 470 | 83.66 | |
| Publication status | | | | | 99.65 | 0.48 |
| Unpublished | 0.29 | 0.05 | 74 | 345 | 54.25 | |
| Published | 0.34 | 0.04 | 117 | 764 | 105.92 | |
| NCEE trial | | | | | 16.19 | 0.43 |
| Not an NCEE trial | 0.33 | 0.03 | 177 | 1048 | 149.15 | |
| NCEE trial | 0.27 | 0.07 | 14 | 61 | 14.42 | |
| Assumed correlation[a] | | | | | 2.81 | 0.39 |
| Not assumed | 0.32 | 0.03 | 189 | 1099 | 152.65 | |
| Assumed | 0.21 | 0.11 | 5 | 10 | 2.79 | |
| Attrition and baseline equivalence | | | | | 190.00[b] | 0.73 |
| Low-attrition RCT | 0.30 | 0.04 | 45 | 255 | 36.88 | |
| Baseline equivalence satisfied | 0.34 | 0.04 | 68 | 292 | 52.82 | |
| Neither standard satisfied | 0.32 | 0.04 | 128 | 562 | 107.06 | |
| Level of random assignment | | | | | 190.00[b] | 0.98 |
| Student | 0.32 | 0.04 | 93 | 547 | 75.59 | |
| Teacher | 0.33 | 0.05 | 67 | 379 | 58.32 | |
| School | 0.33 | 0.08 | 33 | 183 | 24.50 | |

*Note*. The first results column (Mean) reports conditional means, which are the predicted values (Hedges' $g$) from a multivariable, mixed-effects meta-regression model that simultaneously controlled for all the listed moderators (e.g., average effect size for student-level assignment when the other moderators were fixed at their observed means). The $p$ values assess the statistical significance of a single moderator or groups of moderators (but not whether individual conditional means for categorical moderators differed from 0). $m$ = number of studies, $k$ = number of effect sizes, $df$ = degrees of freedom.
[a]Assumed correlation reflects whether a correlation was imputed in calculating the effect size, which applies to three scenarios: (1) standard deviations for pre-post gain scores were reported, which had to be corrected; (2) the effect size was based on an ANCOVA $F$-test statistic, but the model $R^2$ value was not reported; and (3) the effect size was based on an unstandardized regression coefficient and its standard error, but the posttest standard deviation was not reported.
[b]Methodological guidance currently does not exist for computing the RVE-adjusted degrees of freedom for multigroup $F$ tests when using multiple imputation (Pustejovsky, 2017). For this reason, we used $m$ – 1 as the denominator degrees of freedom as a naïve $F$ test, where $m$ is the number of studies.

**Table 6.** Meta-regression results for sample demographics moderators.

| Moderator | *b* | *SE* | *df* | *p* |
|---|---|---|---|---|
| Average grade level* | −0.02 | 0.02 | 25.10 | 0.13 |
| Prop. male* | 0.02 | 0.22 | 2.74 | 0.94 |
| Prop. white* | 0.01 | 0.21 | 3.91 | 0.95 |
| Prop. special education* | 0.02 | 0.06 | 6.12 | 0.77 |
| Prop. English language learner* | −0.00 | 0.08 | 3.61 | 0.96 |
| Prop. economically disadvantaged* | −0.01 | 0.09 | 3.62 | 0.89 |

*Note*. The first results column (*b*) reports regression coefficients for these continuous moderators, controlling for the other listed moderators. This model also controlled for methods moderators (e.g., level of random assignment), which are not listed here (but see Table 5).
*Indicates regression coefficient rather than conditional mean.

methods moderators were retained as covariates, regardless of their statistical significance, in the following substantive UTOS moderator models.

The elements of the "U" block (which included sample grade level, gender, special education status, English learner status, and economic disadvantage) did not significantly explain effect heterogeneity on their own (see Table 6). The "T" block (which included intervention type, training method, length, delivery mechanism, and breadth) had two elements with $p < .10$: intervention type and delivery mechanism (see Table 7). Supplemental time ($g = 0.55$) interventions had larger average effects than curriculum and pedagogical/instructional interventions ($g = 0.33$ and 0.26, respectively), and teacher- and interventionist-delivered interventions (both $g = 0.37$) had larger average effects than technology-delivered interventions ($g = 0.11$). The elements from the "O" block (which included outcome domain, outcome-intervention alignment,[11] and outcome timing) did not significantly explain effect heterogeneity (see Table 8). The only element of the "S" block (which included urbanicity, geographic region, and publication year) that had $p < .10$ was publication decade ($b = -0.14$), indicating that effects from older mathematics intervention studies were larger than effects from more recent studies (see Table 9).

After examining each block independently, controlling for methods confounds, we created a combined MUTOS meta-regression model (see Method sections for details on the model building process). Table 10 shows results from the combined moderator mixed-effects model that included moderators that had $p < .10$ in the intermediate UTOS block models and all methods moderators (regardless of statistical significance). In this combined meta-regression model, four MUTOS moderators were significant at $p < .05$: intervention type, intervention delivery, publication year, and outcome type. Average effects were larger for supplemental time interventions ($g = 0.53$) than curriculum or pedagogical/instruction interventions ($g = 0.34$ and 0.27, respectively); teacher and interventionist delivery ($g = 0.37$ and 0.39, respectively) than technology delivery ($g = 0.12$); and earlier than later publication decades ($b = -0.14$). In addition,

[11]Outcome-intervention alignment was operationalized as the proportion of overlap between the outcome domains covered in the outcome measure and the content covered in the intervention. For example, if a study used an outcome measure that measured number sense and basic operations, but the intervention focused only on number sense, the alignment score would be 0.50. If the intervention had focused on both number sense *and* basic operations, the alignment score would have been 1.0. If the intervention had not focused on either number sense or basic operations, the alignment score would have been 0.0.

**Table 7.** Meta-regression results for intervention moderators.

| Moderator | Mean or *b** | *SE* | *m* | *k* | *df* | *p* |
|---|---|---|---|---|---|---|
| Intervention type | | | | | 190.00[a] | 0.05 |
| Curriculum | 0.33 | 0.04 | 83 | 443 | 63.21 | |
| Pedagogical/Instructional | 0.26 | 0.05 | 85 | 553 | 69.48 | |
| Supplemental | 0.55 | 0.11 | 24 | 113 | 25.97 | |
| Intervention training | | | | | 190.00[a] | 0.44 |
| None or not reported | 0.32 | 0.08 | 67 | 376 | 50.21 | |
| One-time training | 0.26 | 0.06 | 58 | 353 | 47.06 | |
| Infrequent ongoing training | 0.36 | 0.06 | 38 | 193 | 37.90 | |
| Frequent ongoing training | 0.40 | 0.07 | 30 | 187 | 27.84 | |
| Intervention length | | | | | | |
| Number of hours* | −0.00 | 0.00 | – | – | 13.86 | 0.97 |
| Number of weeks* | −0.00 | 0.00 | – | – | 28.58 | 0.21 |
| Intervention delivery | | | | | 190.00[a] | 0.03 |
| Teacher | 0.37 | 0.05 | 110 | 608 | 63.20 | |
| Technology | 0.11 | 0.10 | 65 | 375 | 34.12 | |
| Interventionist | 0.37 | 0.06 | 52 | 380 | 49.86 | |
| Intervention breadth score* | −0.01 | 0.02 | – | – | 15.19 | 0.76 |

*Note*. The first results column (Mean or *b*) reports conditional means for categorical moderators (e.g., intervention type) and regression coefficients for continuous moderators (e.g., number of weeks). The conditional means for categorical moderators are the predicted values (Hedges' *g*) from a multivariable, mixed-effects meta-regression model that simultaneously controlled for all the listed moderators (e.g., average effect size for curriculum interventions when the other moderators were fixed at their observed means). This model also controlled for methods moderators (e.g., level of random assignment), which are not listed here (but see Table 5). The *p* values assess the statistical significance of a single moderator or groups of moderators (but not whether individual conditional means for categorical moderators differed from 0). $m$ = number of studies, $k$ = number of effect sizes, $df$ = degrees of freedom.
*Indicates regression coefficient rather than conditional mean.
[a]Methodological guidance currently does not exist for computing the RVE-adjusted degrees of freedom for multigroup *F* tests when using multiple imputation (Pustejovsky, 2017). For this reason, we used $m - 1$ as the denominator degrees of freedom as a naïve *F* test, where *m* is the number of studies.

**Table 8.** Meta-regression results for outcome moderators.

| Moderator | Mean or *b** | *SE* | *m* | *k* | *df* | *p* |
|---|---|---|---|---|---|---|
| Outcome domain | | | | | 190.00[a] | 0.48 |
| Basic mathematics | 0.37 | 0.04 | 92 | 578 | 66.84 | |
| Rational numbers/Fractions | 0.31 | 0.06 | 40 | 190 | 35.37 | |
| Algebra | 0.24 | 0.05 | 61 | 216 | 50.12 | |
| Geometry | 0.42 | 0.09 | 47 | 164 | 27.03 | |
| Measurement, data, and/or statistics | 0.35 | 0.06 | 45 | 157 | 28.96 | |
| Outcome-intervention alignment* | 0.04 | 0.11 | – | – | 2.44 | 0.76 |
| Outcome timing | | | | | 190.00[a] | 0.98 |
| Midstream during intervention | 0.37 | 0.11 | 14 | 33 | 8.61 | |
| Immediate posttest | 0.33 | 0.03 | 172 | 898 | 135.82 | |
| Follow-up posttest | 0.32 | 0.05 | 42 | 169 | 24.87 | |
| Combination of time periods | 0.29 | 0.36 | 4 | 9 | 1.55 | |

*Note*. The first results column (Mean or *b*) reports conditional means for categorical moderators (e.g., outcome type) and regression coefficient for the continuous moderators (i.e., the outcome-intervention alignment score). The conditional means for categorical moderators are the predicted values (Hedges' *g*) from a multivariable, mixed-effects meta-regression model that simultaneously controlled for all the listed moderators (e.g., average effect size for standardized achievement outcomes when the other moderators were fixed at their observed means). This model also controlled for methods moderators (e.g., level of random assignment), which are not listed here (but see Table 5). The *p* values assess the statistical significance of a single moderator or groups of moderators (but not whether individual conditional means for categorical moderators differed from 0). $m$ = number of studies, $k$ = number of effect sizes, $df$ = degrees of freedom.
*Indicates regression coefficient rather than conditional mean.
[a]Methodological guidance currently does not exist for computing the RVE-adjusted degrees of freedom for multigroup *F* tests when using multiple imputation (Pustejovsky, 2017). For this reason, we used $m - 1$ as the denominator degrees of freedom as a naïve *F* test, where *m* is the number of studies.

**Table 9.** Meta-regression results for setting moderators.

| Moderator | Mean or $b$* | $SE$ | $m$ | $k$ | $df$ | $p$ |
|---|---|---|---|---|---|---|
| Urbanicity | | | | | 190.00[a] | 0.75 |
| Suburban | 0.40 | 0.07 | 51 | 299 | 34.14 | |
| Urban | 0.31 | 0.05 | 82 | 482 | 66.88 | |
| Rural | 0.32 | 0.10 | 39 | 222 | 18.78 | |
| U.S. region | | | | | 190.00[a] | 0.90 |
| West | 0.34 | 0.07 | 35 | 206 | 27.91 | |
| Midwest | 0.26 | 0.09 | 30 | 188 | 19.49 | |
| Southwest | 0.39 | 0.09 | 41 | 260 | 24.31 | |
| Northeast | 0.35 | 0.08 | 51 | 364 | 36.61 | |
| Southeast | 0.29 | 0.06 | 64 | 326 | 36.03 | |
| Publication decade* | −0.14 | 0.07 | – | – | 35.74 | 0.05 |

*Note*. The first results column (Mean or $b$) reports conditional means for categorical moderators (i.e., urbanicity and U.S. region) and regression coefficient for the continuous moderator (i.e., publication year). The conditional means for categorical moderators are the predicted values (Hedges' $g$) from a multivariable, mixed-effects meta-regression model that simultaneously controlled for all the listed moderators (e.g., average effect size for urban samples when the other moderators were fixed at their observed means). This model also controlled for methods moderators (e.g., level of random assignment), which are not listed here (but see Table 5). The $p$ values assess the statistical significance of a single moderator or groups of moderators (but not whether individual conditional means for categorical moderators differed from 0). $m$ = number of studies, $k$ = number of effect sizes, $df$ = degrees of freedom.
*Indicates regression coefficient rather than conditional mean.
[a]Methodological guidance currently does not exist for computing the RVE-adjusted degrees of freedom for multigroup $F$ tests when using multiple imputation (Pustejovsky, 2017). For this reason, we used $m - 1$ as the denominator degrees of freedom as a naïve $F$ test, where $m$ is the number of studies.

researcher-generated measures ($g = 0.45$) continued to yield much larger average effects than standardized measures ($g = 0.15$), by a factor of almost three.

## *Robustness of the MUTOS Model*

We compared the performance of the MUTOS meta-regression model to a machine learning model (random forest) in predicting effect sizes (van Lissa, 2017, 2020). We used leave-one-out cross-validation to compute $R^2$ values to ensure a fair comparison across model type (see the supplemental materials for more detail). In short, the predictions for a study's effect sizes were based on models that did not include that study's effect sizes in the model estimation. We repeated this process for all included studies, yielding two sets of predictions (one based on meta-regression and the other based on machine learning). We then computed $R^2$ values based on the reduction of effect heterogeneity when using these predictions.

Results indicated that both models provided useful predictions of effect sizes, even when using rigorous cross-validation to assess model performance. However, the random forest model (cross-validated $R^2$ = 13%) explained more heterogeneity than the MUTOS meta-regression model (cross-validated $R^2$ = 8%). That is, the machine learning model identified additional information in the coded moderators that helped predict effect sizes (see Figure S2 in the supplemental materials for variable importance rankings).

Regarding specific moderators, the two modeling approaches strongly agreed that the best predictor of effect heterogeneity was outcome type (i.e., larger effect sizes for researcher-generated than standardized measures). The models also agreed that effect sizes were smaller, on average, for technology-delivered interventions than other intervention delivery methods.

**Table 10.** Moderator results from mixed-effects meta-regression model.

| Moderator | Mean or *b** | *SE* | *m* | *k* | *df* | *p* |
|---|---|---|---|---|---|---|
| UTOS moderators | | | | | | |
| Intervention type | | | | | 190.00[b] | 0.04 |
| Curriculum | 0.34 | 0.04 | 83 | 443 | 66.49 | |
| Pedagogical/Instructional | 0.27 | 0.04 | 85 | 553 | 68.52 | |
| Supplemental | 0.53 | 0.10 | 24 | 113 | 24.07 | |
| Intervention delivery | | | | | 190.00[b] | 0.01 |
| Teacher | 0.37 | 0.05 | 110 | 608 | 64.03 | |
| Technology | 0.12 | 0.08 | 65 | 375 | 30.76 | |
| Interventionist | 0.39 | 0.06 | 52 | 380 | 48.11 | |
| Publication decade* | −0.14 | 0.06 | – | – | 36.23 | 0.04 |
| Methods Moderators | | | | | | |
| Outcome type | | | | | 39.80 | <0.01 |
| Researcher-generated measure | 0.45 | 0.05 | 123 | 639 | 101.41 | |
| Standardized achievement measure | 0.15 | 0.05 | 107 | 470 | 76.08 | |
| Publication status | | | | | 87.69 | 0.36 |
| Unpublished | 0.29 | 0.05 | 74 | 345 | 53.33 | |
| Published | 0.34 | 0.04 | 117 | 764 | 99.37 | |
| NCEE trial | | | | | 17.72 | 0.28 |
| Not an NCEE trial | 0.33 | 0.03 | 177 | 1048 | 124.10 | |
| NCEE trial | 0.25 | 0.07 | 14 | 61 | 15.78 | |
| Assumed correlation | | | | | 2.84 | 0.27 |
| Not assumed | 0.33 | 0.03 | 189 | 1099 | 128.34 | |
| Assumed[a] | 0.18 | 0.11 | 5 | 10 | 2.81 | |
| Attrition and baseline equivalence | | | | | 190.00[b] | 0.75 |
| Low-attrition RCT | 0.31 | 0.04 | 45 | 255 | 37.62 | |
| Baseline equivalence satisfied | 0.35 | 0.04 | 68 | 292 | 50.66 | |
| Neither standard satisfied | 0.32 | 0.04 | 128 | 562 | 94.22 | |
| Level of random assignment | | | | | 190.00[b] | 0.74 |
| Student | 0.32 | 0.05 | 93 | 547 | 59.51 | |
| Teacher | 0.31 | 0.05 | 67 | 379 | 62.59 | |
| School | 0.38 | 0.08 | 33 | 183 | 32.05 | |

*Note*. The first results column (Mean or *b*) reports conditional means for categorical moderators (e.g., intervention type) and regression coefficients for continuous moderators (i.e., number of weeks and publication year). The conditional means for categorical moderators are the predicted values (Hedges' *g*) from a multivariable, mixed-effects meta-regression model that simultaneously controlled for all the listed moderators (e.g., average effect size for curriculum interventions when the other moderators were fixed at their observed means). The standard errors (*SE*) were adjusted for effect size dependencies using robust variance estimation. The *p* values assess the statistical significance of a single moderator or groups of moderators (but not whether individual conditional means for categorical moderators differed from 0). *m* = number of studies, *k* = number of effect sizes, *df* = degrees of freedom.
*Indicates regression coefficient rather than conditional mean.
[a]Assumed correlation reflects whether a correlation was imputed in calculating the effect size, which applies to three scenarios: (1) standard deviations for pre-post gain scores were reported, which had to be corrected; (2) the effect size was based on an ANCOVA *F*-test statistic, but the model $R^2$ value was not reported; and (3) the effect size was based on an unstandardized regression coefficient and its standard error, but the posttest standard deviation was not reported.
[b]Methodological guidance currently does not exist for computing the RVE-adjusted degrees of freedom for multigroup *F* tests when using multiple imputation (Pustejovsky, 2017). For this reason, we used $m - 1$ as the denominator degrees of freedom as a naïve *F* test, where *m* is the number of studies.

The two modeling approaches also provided partial agreement on the role of average grade level, though the random forest model suggested additional nuance. Grade level strongly predicted effect sizes in the random forest model (i.e., ranked within the top 3 most important predictors), but the relationship was nonlinear (see Figure S3 in the supplemental materials). Average effects declined from Grades 3 to 7, representing the transition from upper elementary school to middle school. The average effect size was more stable for other grade level ranges. This result provided some agreement with the MUTOS modeling approach, where higher grade levels tended to predict weaker effect sizes ($b = -0.02$) in the intermediate "U – Units" meta-regression model (Table 6). However, this overall linear trend was not statistically significant ($p = .13$) and did not meet our chosen $p = .10$ threshold for inclusion in the combined MUTOS meta-regression model (Table 10). Nevertheless, when considered together, these results overall suggest that average intervention effects may decline in later grade levels, but the relationship may be nonlinear.

The random forest model also suggested additional nuance about intervention length (i.e., the number of weeks students were exposed to the intervention). Similar to grade level, intervention length ranked within the top 3 most important moderators for improving effect size predictions in the random forest model (see Figure S2 in the supplemental materials). Having a medium length of about one half of a school year (~15–20 weeks) predicted the strongest intervention effects, especially for researcher-generated measures (see Figure S3 in the supplemental materials; in contrast, longer interventions generally had weaker effects for standardized measures with no peak in average effects for medium-length interventions). Compared to these medium-length interventions, the model predicted weaker average effects for shorter interventions (e.g., lasting less than one month) and longer interventions (e.g., lasting one full school year). These potential nonlinearities may help explain why intervention length was not retained in the MUTOS model building process (which only tested for an overall linear effect), despite emerging as a key predictor in the random forest model.

The random forest model also suggested caution about the robustness of results for two moderators that were statistically significant in the MUTOS meta-regression model: (a) publication year (decreasing with time) and (b) intervention type (larger for supplemental time interventions). Both were significant in the combined MUTOS meta-regression model and their corresponding intermediate models. However, the random forest model determined that these moderators did not tend to improve effect size predictions; an automated algorithm did not include these moderators in the final random forest model (for details on this variable selection algorithm, see van Lissa, 2020). One explanation for this discrepancy might be potential confounds with other moderators. The random forest model might have captured other nonlinearities or interactions that covaried with publication year and supplemental time interventions, potentially leaving those moderators as no longer useful predictors of effect sizes (despite emerging as important in the MUTOS meta-regression model). For example, supplemental time interventions tend to be more intensive than other interventions, both in terms of the average number of weeks (27 weeks) and average number of hours (60 h) compared to other intervention types (19 weeks and 20 h, respectively). The random forest model might have accounted for these confounds in intervention intensity differently, potentially explaining the diverging results about average differences in effect sizes across intervention type.

### *Selective Reporting Bias Analyses*

Although our literature search aimed to systematically find unpublished studies, our results could nevertheless be influenced by selective reporting (e.g., authors publishing only studies or outcomes with significant effects). Of the 191 RCTs included in our analyses, 39% came from gray literature sources such as doctoral dissertations and conference papers.

As detailed in the supplemental materials, we examined selective reporting bias (i.e., publication bias) by both (a) comparing average effects from published versus unpublished studies and (b) testing and adjusting for small-study effects (e.g., small studies with small observed effects may not be published due to lack of statistical significance). Both approaches yielded similar conclusions. Without controlling for any moderators, they provided some suggestive evidence of selective reporting bias (e.g., smaller average effects for unpublished versus published studies and larger versus smaller studies). For example, unpublished studies had somewhat smaller average effects than published studies ($g = 0.24$ versus 0.36, respectively), which was a statistically significant difference ($b = 0.12$, $SE = 0.06$, $p = .05$).

The evidence supporting selective reporting bias largely disappeared, however, once we adjusted for the MUTOS moderators in Table 10. Confounds other than selective reporting might therefore account for results such as published-unpublished differences. For example, unpublished studies used standardized measures more often than published studies (54% versus 37% of effect sizes, respectively). The larger half of studies (based on a median split in effect size variance) also used standardized measures more often than the smaller half of studies (56% versus 30%, respectively). Differences in using standardized achievement measures might therefore partly account for the somewhat smaller average effects from unpublished studies and larger studies.

The supplemental materials describe these findings in more detail, along with results from another selective reporting analysis method (i.e., selection models; Vevea & Hedges, 1995) and sensitivity analyses that explored the consequences of varying magnitudes of selective reporting bias (Mathur & VanderWeele, 2020; see also Vevea & Woods, 2005). Selection models yielded inconclusive results (i.e., implausible adjusted mean estimates that were highly sensitive to model specifications). The sensitivity analyses nevertheless suggested our meta-analytic estimates of intervention effects were robust to plausible magnitudes of reporting bias. Overall, these selective reporting analyses support confidence in our estimates of average mathematics intervention effects, especially after accounting for potential confounds (such as use of standardized vs. researcher-generated measures in published vs. unpublished studies).

## Discussion

This study took a high-level review of a quarter century worth of experimental research in U.S. PreK-12 mathematics education, focusing on understanding variation in mathematics intervention effects. The results of our review and synthesis indicate that the middle 95% of intervention effects tend to vary between about −0.60 and 1.23 standard deviations. Although the results indicate wide heterogeneity, they also tell an important story about the probability of positively impacting student learning outcomes. As noted in the Results section, the probability that a random mathematics intervention effect has a positive impact is 75%, and the probability that the intervention effect is at least 0.10 and 0.25 standard deviations is 68% and 55%, respectively. This high-level review of effects is an important perspective, especially when researchers and policymakers often infer that "nothing works."

While our study was able to identify several consistent sources of effect heterogeneity, largely from methodological and intervention characteristics, the effect sizes in our synthesis were generally weakly related to theoretically important characteristics of the samples, outcomes, and settings. On the one hand, these results describe a general robustness of mathematics intervention effects for different kinds of learners in different contexts and for different content areas. On the other hand, our analyses explained about 10% of intervention heterogeneity, which may indicate that we were likely unable to observe and systematically code other meaningful study characteristics.

Nevertheless, the explanatory power of our combined MUTOS meta-regression was strong, relative to other large-scale meta-analyses in STEM subject areas. As example comparisons, we downloaded the raw data for two large meta-analyses on science education intervention studies (Taylor et al., 2018) and computer-based scaffolding in STEM education (Belland et al., 2017). We found that the percentage of explained heterogeneity was 2% or less in those meta-analyses, based on using the moderators that the original meta-analysis authors had coded for. In contrast, a more focused meta-analysis on STEM professional development interventions explained a much higher percentage of heterogeneity, 30% or higher, depending on the model (Lynch et al., 2019). However, the unconditional effect heterogeneity was also much smaller in the Lynch et al. (2019) study (0.19 standard deviations) compared to our meta-analysis (0.47 standard deviations), meaning that the same change in absolute heterogeneity will yield a larger change in percentage heterogeneity explained. Interestingly, one of the most explanatory study features that Lynch et al. found was outcome measure type (i.e., researcher-developed vs. standardized), consistent with our results discussed in the following sections.

In the following sections, we connect the results of our synthesis to those of related, and recent, syntheses in STEM education. We further describe the limitations of this study and provide a link to an online application that allows users to download and interact with the study data directly (https://osf.io/f9gud/?view_only=c97ba1316ff44606b8954d686e4d2d8b).

## Methodological Characteristics

The most explanatory study feature that our analyses identified was outcome measure type: whether the measure was researcher-developed or standardized. In our final MUTOS meta-regression model, researcher-developed measures had effects that were 0.30 standard deviations larger on average than standardized measures. Other recent syntheses in STEM education have found similar results (for a broader review of syntheses in education, see Wolf, 2021). Compared to standardized measures, researcher-generated measures yielded larger effect sizes by 0.17 standard deviations for mathematics intelligent tutoring systems (Steenbergen-Hu & Cooper, 2013), 0.26 standard deviations for science education interventions (Taylor et al., 2018), and 0.27 standard deviations for STEM professional development programs (Lynch et al., 2019). These differences remained after controlling for other study features in meta-regression models.

Wolf (2021) discussed various hypotheses for the stronger effects for researcher-generated measures, including differences in narrow versus broad measurement constructs, implementation fidelity, developer conflicts of interest, and reliability and validity. One hypothesis we investigated was whether researcher-developed measures were better aligned to the interventions under investigation (i.e., we coded for the fraction of outcome domains in the outcome measure that were also covered in the intervention; see Footnote 18 for details). We found that the differences in average effect sizes from researcher-generated measures versus standardized measures remained even after adjusting for intervention-outcome alignment. However, our indicator for alignment was based on broad outcome domains such as "algebra" or "geometry," which may have been too coarse to pick up on important finer-grained distinctions (e.g., subtopic in algebra such as linear equations).

Variation in other methodological study features did not strongly correspond to variation in effect magnitude, which is an encouraging indication of methodological robustness. For example, when adjusting for other features, studies with high attrition and baseline imbalance on pretest measures, as defined by the WWC, had roughly comparable average effect estimates as those that had low attrition and demonstrated baseline equivalence on the pretest measures. Similarly, we did not observe substantial differences in effect size across different levels of random assignment (i.e., school, teacher, student). Not adjusting for moderator, published studies had somewhat larger effect sizes than unpublished studies, a finding that emerges in many research syntheses; however, the difference was not statistically significant after adjusting for other features, consistent with our selective reporting bias analyses.

## Sample Characteristics

Results suggested that average effects tended to decline in higher student grade levels. However, the empirical evidence for this decline had some sensitivity to the modeling approach. The strongest evidence came the from the machine learning approach (i.e., random forest model) which automatically modeled nonlinearities and interactions. This model placed student grade level within the top 3 moderators that best improved effect size predictions, with average effects declining most rapidly in the transition from elementary school to middle school (roughly Grades 3–7). The (linear) meta-regression also found a negative trend for grade level ($b = -0.02$, $p = 0.13$), though the coefficient was not statistically significant at $p < .10$, potentially due to unmodeled nonlinearities.

The grade-related trends are consistent with other research indicating that students' mathematics learning may be most malleable in earlier ages. For instance, in business-as-usual instruction, average one-year growth in student mathematics achievement is 1.14 standard deviations from kindergarten to Grade 1, compared to only 0.22 standard deviations from Grades 8 to 9 (Bloom et al., 2008; Table 3). Relatedly, some meta-analyses have found some suggestive evidence for stronger intervention-comparison effects in earlier grade levels (e.g., Cheung & Slavin, 2013, 2016; Nickow et al., 2020), though this difference has not always consistently emerged (e.g., Taylor et al., 2018). For example, one broad review (Cheung & Slavin, 2016) found larger effects in elementary school

($g = 0.20$) than secondary school ($g = 0.17$), though the difference was not significant ($p = .06$). Our results suggest a nuanced relationship in which specific transition points (e.g., elementary to middle school) may be especially important. However, our results should be interpreted cautiously given the exploratory nature of the random forest model. We encourage future primary research and meta-analyses to investigate these differences more thoroughly.

Despite the grade level findings, effect sizes were largely unrelated to other sample characteristics such as study compositions of gender, race, special education status, English learner status, and economically disadvantaged status. This result has at least three possible explanations. First, the combined MUTOS meta-regression model has limited statistical power for individual model coefficients. The estimated variances are adjusted both for effect size dependencies and for distributional violations (e.g., non-normal or imbalanced moderators). Furthermore, rates of missing data were highest for sample demographics, decreasing precision and statistical power for detecting those moderator effects. Second, Cooper and Patall (2009) note that within-study moderators, especially those that are inherent to the individuals in the study samples, do not necessarily translate to study-level moderators, which is known as ecological fallacy. This is an important caveat when interpreting meta-analytic results that rely on aggregations of individual-level characteristics. Third, other characteristics of students, teachers, learning environments, and implementation may have explained additional effect heterogeneity, but these characteristics were unobserved or otherwise uncoded, such as emotional states (e.g., Barroso et al., 2021), school climate (e.g., Kwong & Davis, 2015), and teacher credentials and experience (e.g., Nye et al., 2004).

## Intervention Characteristics

One intervention characteristic that was consistently related to effect magnitude was intervention delivery mechanism. In the combined MUTOS meta-regression model, teacher- and interventionist-delivered programs ($gs = 0.37$ and 0.39, respectively) had average effects that were about three times as large as effects from technology-delivered programs ($g = 0.12$). The random forest model provided converging evidence for this difference. The conditional mean of 0.12 standard deviations for technology-delivered programs is similar to the average effect of 0.15 standard deviations that Cheung and Slavin (2013) found for similar interventions.

The weaker effect for technology-delivered mathematics programs is an important and timely finding given the recent context of the COVID-19 pandemic. In 2020, the pandemic forced most large U.S. school districts to rapidly switch to fully remote education, disrupting the learning of millions of U.S. children (Sahni et al., 2021). One hypothesis for the intervention delivery differences we found (based on studies prior to the pandemic) is that online or virtual instruction may result in less sustained student engagement than interventions involving teachers, aides, or other instructional staff (Blasiman et al., 2018). It is also important to note that technology-delivered programs are a broad, heterogeneous group of interventions, which can include in-person instruction (e.g., teachers instructing students how to use computer learning programs during normal classroom hours). Nevertheless, at a broad level, technology delivery is associated with weaker effects, highlighting the critical importance of understanding what types of technology delivery can work to improve student mathematics achievement.

Two other intervention characteristics that our analyses identified as potentially explanatory were intervention type and intervention length. We describe these as "potentially explanatory" because the results were sensitive to modeling strategy. Regarding intervention type, supplemental time interventions ($g = 0.53$) had larger average effects than curriculum interventions ($g = 0.34$) or instructional/pedagogical interventions ($g = 0.27$), after adjusting for other study features in the combined MUTOS meta-regression model. This result is similar to Cheung and Slavin's (2013) finding that supplemental computer assisted programs had larger average effects than educational technology programs integrated into regular mathematics instruction (see also Nickow et al., 2020, for related evidence on comparatively large effects for tutoring programs; Kraft & Falken, 2021). One explanation for these findings is that supplemental time programs simply provide more opportunities to learn (i.e., more instructional time overall). Another explanation is that supplemental time programs are more intensive and sustained than other programs, as supported by our coded data. Supplemental time interventions lasted approximately 60 h over 27 weeks on average, compared to 24 h over 21 weeks for curriculum interventions and 16 h over 17 weeks for instructional interventions. The evidence for differences in average effects across these intervention types is tentative, however, because the random forest model did not select intervention type as a moderator that improved effect size predictions. However, this result only emerged in the primary MUTOS model; it was not identified as a key moderator in our exploratory random forest model.

Conversely, the combined MUTOS model did not identify intervention length as a key moderator, but the exploratory random forest model did. The difference in results from the meta-regression model may stem from the complex relationship found for intervention length. In the random forest model, average effects were largest for medium-length interventions lasting approximately one-half of a school year (about 15–20 weeks). Researcher-generated measures drove this effect; in contrast, effects for standardized measures tended to very slightly decline throughout with no peak in average effects for medium length. These results contrast with simple, intuitive predictions that longer interventions should typically yield stronger effects. However, the results partially align with other meta-analyses in education that typically find no moderation by length (e.g., Kraft et al., 2018) or weaker effects for longer interventions (e.g., Dietrichson et al., 2017; Nelson & McMaster, 2019; Steenbergen-Hu & Cooper, 2013). Future research should more thoroughly investigate the reasons for these potential differences in effect sizes (e.g., whether researchers' control of the testing environment may explain the larger effects for medium-length interventions; see Cheung & Slavin, 2013).

## Outcome and Settings Characteristics

For outcome characteristics, the magnitude of average effects varied among the outcome domains examined in this review, but the differences were not statistically significant. For example, average effect sizes for algebra measures ($g = 0.24$) were about 0.18

standard deviations smaller than for geometry domain ($g = 0.42$), but the difference was not statistically significant. We also found little evidence of moderation for the timing of outcome assessment or the level of alignment between outcomes and intervention content. However, one key limitation is that only 15% of effect sizes were for follow-up outcomes measured with some delay after the end of the intervention, offering limited evidence on the longevity of the intervention effects.

For setting characteristics, the combined MUTOS meta-regression suggested that effect sizes declined by 0.14 standard deviations on average for each additional decade. A meta-analysis of K-12 technology-enhanced mathematics interventions found a similar trend across decades, though the differences were not significant (Cheung & Slavin, 2013; Table 4). One explanation for this result is increased methodological rigor and scientific standards for causal inference research in mathematics education. That is, changes in the cultural norms of scientific practice may help address previously unmitigated sources of bias. Another explanation is that students are less responsive to intervention than they were in earlier decades. For example, as more innovative programs enter the mainstream, the counterfactual (or "business as usual") might become increasingly competitive with experimental programs. Another contributor might be changes in mathematical standards and tests such as the introduction of the No Child Left Behind Act in the early 2000s. However, this result was also not robust to modeling strategy as the random forest model did not identify publication year as an important moderator, adjusting for other study characteristics. As such, we caution against overinterpretation.

## Limitations

Our review included 191 randomized controlled trials that covered a quarter century of PreK-12 mathematics intervention research, along with systematically coding for methods, sample, intervention, outcome, and setting characteristics. The large scale of our review was a key strength for exploring effect heterogeneity, offering greater statistical power for moderator analyses than in smaller-scale reviews focusing on specific types of mathematics interventions or subpopulations. The scale, however, also presented practical constraints. One of those constraints was being unable to code for more granular study characteristics that might further explain effect heterogeneity, as one might in a more focused or traditional meta-analysis. Research reporting was also a key challenge. For example, most studies did not report how many special education, English learner, or economically disadvantaged students were in their samples, limiting our ability to understand broad relationships between mathematics intervention effects and sample characteristics.

The comprehensiveness of study reporting practices was another limitation we encountered. While the collective set of moderators used in our analyses explained about 10% of the heterogeneity in mathematics intervention effects, we suspect that most variation in intervention effects is likely due to idiosyncratic characteristics of the studies that are rarely reported in a systematic way. For example, the counterfactual conditions (operationalized as "business-as-usual" in this review) likely widely varied across studies, but researchers rarely systematically measure and report a service contrast in a way that is useful for meta-analysis. Also, although we coded for researcher-reported dispositions about implementation fidelity (i.e., whether the researchers indicated the intervention was delivered as intended), there is likely far more granular variation in effective implementation than what we were able to systematically code for (e.g., idiosyncratic study-specific aspects of implementation that cannot be easily compared across studies in meta-analyses).

Alternative approaches to coding could also uncover further nuance. For example, we chose to identify a primary intervention type for each study (i.e., curriculum, instructional, or supplemental time), rather than having a check-all-that-apply intervention typology. Identifying the primary type was not always immediately clear because some interventions had multiple, relatively balanced, components. Lynch et al. (2019), for example, coded for overlapping professional development intervention components and found evidence of a small positive relationship with effect sizes, although the overall results of their study are very similar to the results we found for pedagogical and instructional interventions in this study.

We evaluated the robustness of our results by including two modeling approaches—one based on a planned meta-regression using the MUTOS framework and the other data-driven approach based on machine learning using automated algorithms. Including both approaches strengthened our analyses, but the contrast revealed potential limitations in the standard linearity assumptions for the MUTOS meta-regression approach. Some moderators such as grade level or intervention length may not have simple linear relationships with intervention effects. Using data-driven modeling strategies like the one we used in this study provide new opportunities for researchers and meta-analysts to evaluate the robustness of their planned modeling strategies (including linearity assumptions), especially in large, complex, reviews like this one.

Though the large scale of our review was a strength, power to detect moderator effects in meta-analysis tends to be low, even with many included studies (Hempel et al., 2013, Hedges & Pigott, 2001). For this reason, we still presented conditional mean effect sizes for each of moderators we examined (regardless of statistical significance) as there is value in understanding the patterns of average effects across their levels, even if they are estimated with varying degrees of precision. Failure to detect statistically significant moderator effects should not be interpreted as strong evidence of no moderation, especially for characteristics such as racial and socioeconomic demographics that had major limitations due to missing study-reported information.

Last, we restricted our review to randomized experiments without known confounds, strengthening the internal validity of the included evidence. But there is also a broader evidence base of well-executed quasi-experimental designs that we did not capture. Though quasi-experimental studies tend to produce larger effects than similar RCTs and are more susceptible to selection bias (Cheung & Slavin, 2016), they can offer important evidence for certain types of interventions, learners, or settings that may be underrepresented in the studies we included.

## Conclusions, Future Directions, and Open Science

A quarter century worth of experimental evidence shows that mathematics interventions in U.S. PreK-12 education improve student learning across a wide range of program types, student demographics, and outcome domains. However, these intervention effects also widely vary. Our analyses identified how specific aspects of the study characteristics may help account for this heterogeneity, but much of the heterogeneity remains unexplained based on readily codable information in the study reports. To help others further explore, we are sharing our coded data, codebook, and an interactive web application with the public (https://airshinyapps.shinyapps.io/math_meta_database/). Our team took one set of approaches to organizing and analyzing the evidence, and we hope that this dataset may help answer a host of other questions that researchers and decisionmakers may have. The web application allows users to explore the pool of mathematics intervention effects using evidence gap maps, data visualizations, and opportunities to construct customized meta-analyses. Users may also download the data to use as they see fit. The application and dataset will be maintained and updated periodically by the Methods of Synthesis and Integration Center (MOSAIC) at the American Institutes for Research (https://www.air.org/centers/mosaic).

## Funding

This study was supported by the U.S. Institute of Education Sciences (IES) under grant [R305A170146]. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily represent the views of the IES.

## References

*Study included in meta-analysis

*Abrams, L. S. (2008). *The effect of computer mathematics games on elementary and middle school students' mathematics motivation and achievement* [Unpublished doctoral dissertation]. Capella University, Minneapolis, MN.

Achieve. (2014). *Rising to the challenge*. Author. http://www.achieve.org/rising-challenge

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82*(4), 436–476. https://doi.org/10.3102/0034654312458162.

*Aladé, F., Lauricella, A. R., Beaudoin-Ryan, L., & Wartella, E. (2016). Measuring with Murray: Touchscreen technology and preschoolers' STEM learning. *Computers in Human Behavior*, *62*, 433–441. https://doi.org/10.1016/j.chb.2016.03.080

*Albright, R. E. (2012). *The impact of music on student achievement in the third and fifth grade math curriculum* [Unpublished doctoral dissertation]. Northcentral University, San Diego, CA.

Aloe, A., & Becker, B. J. (2009, April). *Modeling heterogeneity in meta-analysis: Generalizing using Cronbach's (M)UTOS framework and meta-analytic data* [Paper presentation]. Paper presented at the annual meeting of the American Educational Research Association.

*Arnold, D. H., Fisher, P. H., Doctoroff, G. L., & Dobbs, J. (2002). Accelerating math development in Head Start classrooms. *Journal of Educational Psychology*, *94*(4), 762–770. https://doi.org/10.1037/0022-0663.94.4.762

*Arroyo, I., Royer, J. M., & Woolf, B. P. (2011). Using an intelligent tutor and math fluency training to improve math performance. *International Journal of Artificial Intelligence in Education*, *21*(1–2), 135–152. https://doi.org/10.3233/JAI-2011-020

*Axtell, P. K., McCallum, R. S., Bell, S. M., & Poncy, B. (2009). Developing math automaticity using a classwide fluency building procedure for middle school students: A preliminary study. *Psychology in the Schools*, *46*(6), 526–538. https://doi.org/10.1002/pits.20395

*Bai, H., Pan, W., Hirumi, A., & Kebritchi, M. (2012). Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. *British Journal of Educational Technology*, *43*(6), 993–1003. https://doi.org/10.1111/j.1467-8535.2011.01269.x

Barnard, J., & Rubin, D. B. (1999). Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*(4), 948–955. https://doi.org/10.1093/biomet/86.4.948

*Barnes, M. A., Klein, A., Swank, P., Starkey, P., McCandliss, B., Flynn, K., Zucker, T., Huang, C.-W., Fall, A.-M., & Roberts, G. (2016). Effects of tutorial interventions in mathematics and attention for low-performing preschool children. *Journal of Research on Educational Effectiveness*, *9*(4), 577–606. https://doi.org/10.1080/19345747.2016.1191575

Barroso, C., Ganley, C. M., McGraw, A. L., Geer, E. A., Hart, S. A., & Daucourt, M. C. (2021). A meta-analysis of the relation between math anxiety and math achievement. *Psychological Bulletin*, *147*(2), 134–168. https://doi.org/10.1037/bul0000307

*Barrow, L., Markman, L., & Rouse, C. E. (2009). Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy*, *1*(1), 52–74. https://doi.org/10.1257/pol.1.1.52

Basista, B., & Mathews, S. (2002). Integrated science and mathematics professional development programs. *School Science and Mathematics*, *102*(7), 359–370. https://doi.org/10.1111/j.1949-8594.2002.tb18219.x

Becker, B. J. (2017). Improving the design and use of meta-analyses of career interventions. In J. P. Sampson, E. Bullock-Yowell, V. C. Dozier, D. S. Osborn, & J. G. Lenz (Eds.), *Integrating theory, research, and practice in vocational psychology: Current status and future directions* (pp. 95–107). Florida State University.

Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education. *Review of Educational Research*, *87*(2), 309–344. https://doi.org/10.3102/0034654316670999

*Black, A. R., Doolittle, F., Zhu, P., Unterman, R., & Grossman, J. B. (2008). *The evaluation of enhanced academic instruction in after-school programs: Findings after the first year of implementation* (NCEE 2008-4021). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

*Black, A. R., Somers, M.-A., Doolittle, F., Unterman, R., & Grossman, J. B. (2009). *The evaluation of enhanced academic instruction in after-school programs: Final report* (NCEE 2009-4077). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Blasiman, R. N., Larabee, D., & Fabry, D. (2018). Distracted students: A comparison of multiple types of distractions on learning in online lectures. *Scholarship of Teaching and Learning in Psychology*, *4*(4), 222–230. https://doi.org/10.1037/stl0000122

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328. https://doi.org/10.1080/19345740802400072

*Bond, J. B., & Ellis, A. K. (2013). The effects of metacognitive reflective assessment on fifth and sixth graders' mathematics achievement. *School Science and Mathematics*, *113*(5), 227–234. https://doi.org/10.1111/ssm.12021

*Booth, J. L., Oyer, M. H., Pare-Blagoev, A. J., Elliot, A., Barbieri, C., Augustine, A., & Koedinger, K. R. (2015). Learning algebra by example in real-world classrooms. *Journal of Research on Educational Effectiveness*, *8*(4), 530–551. https://doi.org/10.1080/19345747.2015.1055636

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis* (1st ed.). Wiley.

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

*Boster, F. J., Meyer, G. S., Roberto, A. J., Lindsey, L., Smith, R., Inge, C., & Strom, R. E. (2007). The impact of video streaming on mathematics performance. *Communication Education*, *56*(2), 134–144. https://doi.org/10.1080/03634520601071801

*Bottge, B. A., Grant, T. S., Stephens, A. C., & Rueda, E. (2010). Advancing the math skills of middle school students in technology education classrooms. *NASSP Bulletin*, *94*(2), 81–106. https://doi.org/10.1177/0192636510379902

*Bottge, B. A., Heinrichs, M., Mehta, Z. D., & Hung, Y.-H. (2002). Weighing the benefits of anchored math instruction for students with disabilities in general education classes. *The Journal of Special Education*, *35*(4), 186–200. https://doi.org/10.1177/002246690203500401

*Bottge, B. A., Ma, X., Gassaway, L., Toland, M. D., Butler, M., & Cho, S. J. (2014). Effects of blended instructional models on math performance. *Exceptional Children*, *80*(4), 423–437. https://doi.org/10.1177/0014402914527240

*Bottge, B. A., Rueda, E., LaRoque, P. T., Serlin, R. C., & Kwon, J. (2007). Integrating reform-oriented math instruction in special education settings. *Learning Disabilities Research & Practice*, *22*(2), 96–109. https://doi.org/10.1111/j.1540-5826.2007.00234.x

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. (2008). Mathematics intervention for first- and second-grade students with mathematics difficulties: The effects of Tier 2 intervention delivered as booster lessons. *Remedial and Special Education*, *29*(1), 20–32. https://doi.org/10.1177/0741932507309712

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

*Burghardt, M. D., Lauckhardt, J., Kennedy, M., Hecht, D., & McHugh, L. (2015). The effects of a mathematics infusion curriculum on middle school student mathematics achievement. *School Science and Mathematics*, *115*(5), 204–215. https://doi.org/10.1111/ssm.12123

*Cabalo, J. V., Ma, B., & Jaciw, A. (2007). *Comparative effectiveness of Carnegie Learning's Cognitive Tutor Bridge to Algebra curriculum: A report of a randomized experiment in the Maui School District.* Empirical Education Inc.

Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance (Working Paper 14194).* National Bureau of Economic Research. http://www.nber.org/papers/w14194.pdf

*Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*(3), 430–454. https://doi.org/10.1086/657654

*Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts* (NCEE 2009-4041). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Canty, A., & Ripley, B. (2020). *boot: Bootstrap R (S-Plus) functions. R Package Version*, *1*, 3–25.

*Cardelle-Elawar, M. (1995). Effects of metacognitive instruction on low achievers in mathematics problems. *Teaching and Teacher Education*, *11*(1), 81–95. https://doi.org/10.1016/0742-051X(94)00019-3

*Carr, M., Taasoobshirazi, G., Stroud, R., & Royer, J. M. (2011). Combined fluency and cognitive strategies instruction improves mathematics achievement in early elementary school. *Contemporary Educational Psychology*, *36*(4), 323–333. https://doi.org/10.1016/j.cedpsych.2011.04.002

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

*Cary, M. S., Doabler, C., Clarke, B., Fien, H., Baker, S. K., & Jungjohann, K. J. (2013, March). *Evaluating the promise of the FUSION Tier 2 mathematics intervention* [Paper presentation].

Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

*Cates, G. L., & Skinner, C. H. (2000). Getting remedial mathematics students to prefer homework with 20% and 40% more problems: An investigation of the strength of the interspersing procedure. *Psychology in the Schools*, *37*(4), 339–347.

*Cavalluzzo, L., Geraghty, T. M., Steele, J. L., & Alexander, J. K. (2013, September). *Using data to inform decisions: How teachers use data to inform practice and improve student performance in mathematics* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*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 (IRM-2013-U-006508).* CNA Corporation.

*Cavalluzzo, L., Lowther, D., Mokher, C., & Fan, X. (2012). *Effects of the Kentucky Virtual Schools' hybrid program for Algebra I on Grade 9 student math achievement* (NCEE 2012-4020). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/edlabs/regions/appalachia/pdf/20124020.pdf

Change the Equation. (2015). *Solving the diversity dilemma: Changing the face of the STEM workforce.* Author. http://changetheequation.org/sites/default/files/2015%20Solving%20the%20Diversity%20Dilemma%20FINAL%206.2015.pdf

*Chase, K. N. P. (2016). *Building algebra one Giant Step at a time: Toward a reverse-scaffolding pedagogical approach for fostering subjective transparency through engineering levels of interaction with a technological learning environment* [Doctoral dissertation]. https://escholarship.org/uc/item/3j00n1v0

*Cheng, Y., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development*, *15*(1), 2–11. https://doi.org/10.1080/15248372.2012.725186

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

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

*Cho, S., Yang, J., & Mandracchia, M. (2015). Effects of M3 curriculum on mathematics and English proficiency achievement of mathematically promising English language learners. *Journal of Advanced Academics*, *26*(2), 112–142. https://doi.org/10.1177/1932202X15577205

*Clark, C. T. (2012). The development of embodied representations of numerical understanding through gameplay [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 3532452). http://search.proquest.com/docview/1163706162

*Clark, T. F., Arens, S. A., & Stewart, J. (2015). March). *Efficacy study of a pre-algebra supplemental program in rural Mississippi: Preliminary findings* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

*Clarke, B., Baker, S., Smolkowski, K., Doabler, C., Cary, M. S., & 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., Cary, M. S., 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), 160–177. https://doi.org/10.1080/02796015.2014.12087442

*Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*(2), 443–494. https://doi.org/10.3102/0002831207312908

*Clements, D. H., Sarama, J., Layzer, C., Unlu, F., Wolfe, C. B., & Spitler, M. E. (2013, March). *Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and*

*technologies: Persistence of effects three years after treatment* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Clements, D. H., Sarama, J., Layzer, C., Unlu, F., Wolfe, C. B., Spitler, M. E., & Weiss, D. (2016). March). *Effects of TRIAD on mathematics achievement: Long-term impact* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127–166. https://doi.org/10.5951/jresematheduc.42.2.0127
*Clements, D. H., Sarama, J., Spitler, M. E., & Wolfe, C. B. (2011, March). *Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Mechanisms of persistence of effects* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812–850. https://doi.org/10.3102/0002831212469270
*Codding, R. S., VanDerHeyden, A. M., Martin, R. J., Desai, S., Allard, N., & Perrault, L. (2016). Manipulating treatment dose: Evaluating the frequency of a small group intervention targeting whole number operations. *Learning Disabilities Research & Practice*, *31*(4), 208–220. https://doi.org/10.1111/ldrp.12120
*Cole, C. A. (2008). *Ten weeks of academic intervention designed to improve math word problem solving among middle school students: Effects of a randomized pilot study* [Unpublished doctoral dissertation]. University of South Carolina, Columbia, SC.
Coles, C. D., Kable, J. A., & Taddeo, E. (2009). Math performance and behavior problems in children affected by prenatal alcohol exposure: Intervention and follow-up. *Journal of Developmental and Behavioral Pediatrics*, *30*(1), 7–15.
Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165–176. https://doi.org/10.1037/a0015565
*Courtright, C. A. (2017). *Integrating reading into math instruction to increase academic achievement of English language learners* [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 10179211).
*Cramer, K. A., Post, T. R., & Delmas, R. C. (2002). Initial fraction learning by fourth-and fifth-grade students: A comparison of the effects of using commercial curricula with the effects of using the rational number project curriculum. *Journal for Research in Mathematics Education*, *33*(2), 111–144. https://doi.org/10.2307/749646
*Cressey, J., & Ezbicki, K. (2008). *Improving automaticity with basic addition facts: Do taped problems work faster than cover, copy, compare?* NERA Conference Proceedings 2008, 12. http://opencommons.uconn.edu/cgi/viewcontent.cgi?article=1003&context=nera_2008
Cronbach, L. J. (1982). *Designing evaluations of educational and social programs* (1st ed.). Jossey-Bass.
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., Baker, S. K., Smolkowski, K., & Fien, H. (2016). Effects of a core kindergarten mathematics curriculum on the mathematics achievement of Spanish-speaking English learners. *School Psychology Review*, *45*(3), 343–361. https://doi.org/10.17105/SPR45-3.343-361
*Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano , Means , Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first*

*student cohort* (NCEE 2007-4005). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
*Dyson, N., Jordan, N. C., Beliakoff, A., & Hassinger-Das, B. (2015). A kindergarten number-sense intervention with contrasting practice conditions for low-achieving children. *Journal for Research in Mathematics Education*, *46*(3), 331–370. https://doi.org/10.5951/jresematheduc.46.3.0331
*Dyson, N. I. (2011). *A number sense intervention for urban kindergartners at risk for mathematics difficulties* [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 3465744).
*Eno, J., & Heppen, J. (2014). September). *Targeting summer credit recovery: Heterogeneity of treatment effects and gaps between credit recovery students and "on track" students over time* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.
*eSchool News (2004, August 1). *Classroom instruction: Video on demand boosts students' math scores*. eSchool News.
*Falco, L. D. (2008). *"Skill-builders": Enhancing middle school students' self-efficacy and adaptive learning strategies in mathematics* [Unpublished doctoral dissertation]. The University of Arizona, Tucson, AZ.
Fayer, S., Lacey, A., & Watson, A. (2017). *STEM occupations: Past, present, and future*. U.S. Bureau of Labor Statistics. https://www.bls.gov/spotlight/2017/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future/pdf/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future.pdf
*Fede, J. L., Pierce, M. E., Matthews, W. J., & Wells, C. S. (2013). The effects of a computer-assisted, schema-based instruction intervention on word problem-solving skills of low-performing fifth grade students. *Journal of Special Education Technology*, *28*(1), 9–21. https://doi.org/10.1177/016264341302800102
*Fien, H., Doabler, C. T., Nelson, N. J., Kosty, D. B., Clarke, B., & Baker, S. K. (2016). An examination of the promise of the NumberShire level 1 gaming intervention for improving student mathematics outcomes. *Journal of Research on Educational Effectiveness*, *9*(4), 635–661. https://doi.org/10.1080/19345747.2015.1119229
*Fisher, K. R. (2010). *Exploring the mechanisms of guided play in preschoolers' developing geometric shape concepts* [Unpublished doctoral dissertation]. Temple University, Philadelphia, PA.
*Fede, J. L. (2010). *The effects of GO Solve Word Problems mathematics intervention on applied problem solving skills of low performing fifth grade students* [Doctoral dissertation]. http://scholarworks.umass.edu/open_access_dissertations/236
*Foreman, J. L., & Gubbins, E. J. (2015). Teachers see what ability scores cannot: Predicting student performance with challenging mathematics. *Journal of Advanced Academics*, *26*(1), 5–23. https://doi.org/10.1177/1932202X14552279
Frykholm, J. A., & Meyer, M. (2002). Integrated instruction: Is it science? *Is It Mathematics? Mathematics Teaching in Middle School*, *7*(9), 502–508.
*Fuchs, D., Fuchs, L. S., & Fernstrom, P. (1993). A conservative approach to special education reform: Mainstreaming through transenvironmental programming and curriculum-based measurement. *American Educational Research Journal*, *30*(1), 149–177. https://doi.org/10.3102/00028312030001149
*Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, *97*(3), 493–513. https://doi.org/10.1037/0022-0663.97.3.493
*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, *28*(3), 617–641. https://doi.org/10.3102/00028312028003617
*Fuchs, L. S., Fuchs, D., & Karns, K. (2001). Enhancing kindergartners' mathematical development: Effects of peer-assisted learning strategies. *The Elementary School Journal*, *101*(5), 495–510. https://doi.org/10.1086/499684

*Fuchs, L. S., Fuchs, D., & Prentice, K. (2004). Responsiveness to mathematical problem-solving instruction: Comparing students at risk of mathematics disability with and without risk of reading disability. *Journal of Learning Disabilities*, *37*(4), 293–306.

*Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., & Schroeter, K. (2003). Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology*, *95*(2), 306–315. https://doi.org/10.1037/0022-0663.95.2.306

*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., Newman-Gonchar, R., Schumacher, R., Dougherty, B., Bucka, N., Karp, K. S., & Morgan, S. (2021). *Assisting students struggling with mathematics: Intervention in the elementary grades* (WWC 2021-006). National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Science, U.S. Department of Education. https://eric.ed.gov/?id=ED611018

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

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

*Fuchs, L. S., Zumeta, R. O., Schumacher, R. F., Powell, S. R., Seethaler, P. M., Hamlett, C. L., & Fuchs, D. (2010). The effects of schema-broadening instruction on second graders' word-problem performance and their ability to represent word problems with algebraic equations: A randomized control study. *The Elementary School Journal*, *110*(4), 440–463. https://doi.org/10.1086/651191

*Fyfe, E. R. (2016). Providing feedback on computer-based algebra homework in middle-school classrooms. *Computers in Human Behavior*, *63*, 568–574. https://doi.org/10.1016/j.chb.2016.05.082

*Fyfe, E. R., & Rittle-Johnson, B. (2016). The benefits of computer-generated feedback for mathematics problem solving. *Journal of Experimental Child Psychology*, *147*, 140–151.

*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, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://files.eric.ed.gov/fulltext/ED569154.pdf

*Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolittle, F., & Warner, E. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011-4024). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

*Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolittle, F., & Warner, E. (2010). *Middle school mathematics professional development impact study: Findings after the first year of implementation* (NCEE 2010-4009). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

*Gavin, M. K., Casa, T. M., Firmender, J. M., & Carroll, S. R. (2013). The impact of advanced geometry and measurement curriculum units on the mathematics achievement of first-grade students. *Gifted Child Quarterly*, *57*(2), 71–84. https://doi.org/10.1177/0016986213479564

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*(3), 1202–1242. https://doi.org/10.3102/0034654309334431

*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

*Green, K. B. (2014). *The effects of the integration of mathematics within children's literature on early numeracy skills of young children with disabilities* [Doctoral dissertation]. http://scholarworks.gsu.edu/epse_diss/93

*Griesser, S. A. (2001). *A study of the problem solving abilities of seventh grade students who receive anchored problem solving instruction* [Master's thesis]. https://files.eric.ed.gov/fulltext/ED456040.pdf

*Gubbins, E. J., McCoach, D. B., Foreman, J. L., Gilson, C. M., Bruce-Davis, M. N., Rubenstein, L. D., Savino, J., Rambo, K, & Waterman, C. (2013). *What works in gifted education mathematics study: Impact of pre-differentiated and enriched curricula on general education teachers and their students* (RM13242). National Research Center on the Gifted and Talented.

*Harger, J. T. (2007). *An investigation of fourth graders' conceptual understanding of and procedural fluency with rational numbers* [Unpublished doctoral dissertation]. University of Missouri, St. Louis, MO.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

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. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380. https://doi.org/10.3102/1076998610376617

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*(4), 426–445.

*Hegedus, S. J., Dalton, S., & Tapper, J. R. (2015). The impact of technology-enhanced curriculum on learning advanced algebra in U.S. high school classrooms. *Educational Technology Research and Development*, *63*(2), 203–228. https://doi.org/10.1007/s11423-015-9371-z

*Hegedus, S. J., Tapper, J., & Dalton, S. (2016). Exploring how teacher-related factors relate to student achievement in learning advanced algebra in technology-enhanced classrooms. *Journal of Mathematics Teacher Education*, *19*(1), 7–32. https://doi.org/10.1007/s10857-014-9292-5

*Heller, J. I., Curtis, D. A., Rabe-Hesketh, S., & Verboncoeur, C. J. (2007). *The effects of Math Pathways and Pitfalls on students' mathematics achievement: National Science Foundation final report*. National Science Foundation. https://files.eric.ed.gov/fulltext/ED498258.pdf

Hempel, S., Miles, J. N., & Booth, M. J. (2013). Risk of bias: A simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic Review*, *2*, 107. https://doi.org/10.1186/2046-4053-2-107

*Henderson, R. W., & Landesman, E. M. (1995). Effects of thematically integrated mathematics instruction on students of Mexican descent. *The Journal of Educational Research*, *88*(5), 290–300. https://doi.org/10.1080/00220671.1995.9941313

*Heppen, J. (2012). *Broadening access to Algebra I: The impact on eight graders taking an online course*. American Institutes for Research.

*Heppen, J., Allensworth, E., Walters, K., Pareja, A. S., Kurki, A., Nomi, T., & Sorensen, N. (2012). March). *Efficacy of online Algebra I for credit recovery for at-risk ninth-grade students: Evidence from Year 1* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Heppen, J., & Sorensen, N. (2014, September). *Study design and impact results* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Heppen, J. B., Sorensen, N., Allensworth, E., Walters, K., Rickles, J., Taylor, S. S., & Michelman, V. (2017). The struggle to pass algebra: Online versus face-to-face credit recovery for at-risk urban students. *Journal of Research on Educational Effectiveness*, *10*(2), 272–296. https://doi.org/10.1080/19345747.2016.1168500

*Heppen, J. B., Walters, K., Clements, M., Faria, A. M., Tobey, C., Sorensen, N., & Culp, K. (2011). *Access to Algebra I: The effects of online mathematics for Grade 8 students* (NCEE 2012-4021). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

*Hinojosa, T., Miller, S., Swanlund, A., Hallberg, K., Brown, M., & O'Brien, B. (2009). *The Stock Market Game*$^{TM}$ *study: Final report.* Learning Point Associates. https://www.stockmarketgame.org/assets/pdf/2009%20Learning%20Point%20Study%20Full%20Report.pdf

*Hinojosa, T., Miller, S., Swanlund, A., Hallberg, K., Brown, M., & O'Brien, B. (2019). March). *The impact of The Stock Market Game on financial literacy and mathematics achievement: Results from a national randomized controlled trial* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Hohn, R. L., & Frey, B. (2002). Heuristic training and performance in elementary mathematical problem solving. *The Journal of Educational Research*, *95*(6), 374–380. https://doi.org/10.1080/00220670209596612

*Hoshiko, B., Jaciw, A., Ma, B., Miller, G. I., & Wei, X. (2007). *Comparative effectiveness of the Texas Instruments TI-NavigatorTM] Year 2 report of randomized experiments in the East Side Union High School and San Diego Unified School District* (Research Report). Empirical Education Inc.

*Hua, Y., Woods-Groves, S., Kaldenberg, E. R., Lucas, K. G., & Therrien, W. J. (2015). Effects of the TIP strategy on problem solving skills of young adults with intellectual disability. *Education and Training in Autism and Developmental Disabilities*, *50*(1), 31–42.

*Hunt, J. H. (2011). *The effects of a ratio-based teaching sequence on performance in fraction equivalency for students with mathematics disabilities.* Electronic Theses and Dissertations, 2004–2019 (1941). https://stars.library.ucf.edu/etd/1941

*Hutchings, L., Young, C., Almasude, A., McCuaig, S., & Bishara, M. (2006). The rapid acceleration of basic mathematical skills in disadvantaged children: Implementing and assessing an alternative multiplication algorithm. *International Journal of Learning*, *12*(10), 308–317.

*Ikemoto, G. S., Steele, J. L., & Pane, J. E. (2016). Poor implementation of learner-centered practices: A cautionary tale. *Teachers College Record: The Voice of Scholarship in Education*, *118*(13), 1–34. https://doi.org/10.1177/016146811611801309

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, *6*(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

*Irving, K. E., Pape, S. J., Owens, D. T., Abrahamson, L., Silver, D., & Sanalan, V. A. (2016). Classroom connectivity and Algebra I achievement: A three-year longitudinal study. *Journal of Computers in Mathematics and Science Teaching*, *35*(2), 131–151.

*Iseman, J. S., & Naglieri, J. A. (2011). A cognitive strategy instruction to improve math calculation for children with ADHD and LD: A randomized controlled study. *Journal of Learning Disabilities*, *44*(2), 184–195.

*Jabagchourian, J. J. (2008). *Incremental changes in children's multi-digit number representations and arithmetic procedures: Linking strategies and concepts in early mathematics* [Unpublished doctoral dissertation]. University of California at Santa Barbara, Santa Barbara, CA.

*Jaciw, A. P., Cabalo, J. V., & Vu, M. (2007). *Comparative effectiveness of Carnegie Learning's Cognitive Tutor Algebra I curriculum: A report of a randomized experiment in the Maui School District.* Empirical Education Inc.

*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
*Jaciw, A. P., Hegseth, W., & Toby, M. (2015, March). *Assessing impacts of Math in Focus, a "Singapore Math" program for American schools* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.
*Jaciw, A. P., Toby, M., & Ma, B. (2012, September). *Conditions for the effectiveness of a tablet-based algebra program* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Jaciw, A., Toby, M., Ma, B., Lai, G., & Lin, L. (2012). *Measuring the average impact of an iPad algebra program: A report of findings from an RCT in four school districts considering one as a special case* (Empirical Education Rep. No. Empirical_Fuse120330-FR1-YR1_O.3). Empirical Education Inc. https://www.empiricaleducation.com/pdfs/FuseFR.pdf
*Jackson, C. K., & Makarin, A. (2017). *Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment* (Paper No. 22398). National Bureau of Economic Research. https://www.nber.org/papers/w22398.pdf
*Jacob, R. H., 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
*Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education*, *38*(3), 258–288. https://doi.org/10.2307/30034868
*Jayanthi, M., Gersten, R., Taylor, M. J., Smolkowski, K., & Dimino, J. (2017). *Impact of the Developing Mathematical Ideas professional development program on Grade 4 students' and teachers' understanding of fractions* (REL 2017–256). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. http://ies.ed.gov/ncee/edlabs
*Jennings, C. M., Jennings, J. E., Richey, J., & Dixon-Krauss, L. (1992). Increasing interest and achievement in mathematics through children's literature. *Early Childhood Research Quarterly*, *7*(2), 263–276. https://doi.org/10.1016/0885-2006(92)90008-M
*Jiang, Z., White, A., & Rosenwasser, A. (2011). Randomized control trials on the dynamic geometry approach. *Journal of Mathematics Education at Teachers College*, *2*, 8–16.
*Jitendra, A. K., Dupuis, D. N., Rodriguez, M. C., Zaslofsky, A. F., Slater, S., Cozine-Corroy, K., & Church, C. (2013). A randomized controlled trial of the impact of schema-based instruction on mathematical outcomes for third-grade students with mathematics difficulties. *The Elementary School Journal*, *114*(2), 252–276. https://doi.org/10.1086/673199
*Jitendra, A. K., Dupuis, D. N., Star, J. R., & Rodriguez, M. C. (2016). The effects of schema-based instruction on the proportional thinking of students with mathematics difficulties with and without reading difficulties. *Journal of Learning Disabilities*, *49*(4), 354–367.
*Jitendra, A. K., Harwell, M. R., Karl, S. R., Slater, S. C., Simonson, G. R., & Nelson, G. (2016, March). *A replication study to evaluate the effects of schema-based instruction on middle school students' proportional problem-solving performance* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Jitendra, A. K., Star, J. R., Dupuis, D. N., & Rodriguez, M. C. (2013). Effectiveness of schema-based instruction for improving seventh-grade students' proportional reasoning: A randomized experiment. *Journal of Research on Educational Effectiveness*, *6*(2), 114–136. https://doi.org/10.1080/19345747.2012.725804
*Jitendra, A. K., Star, J. R., Rodriguez, M., Lindell, M., & Someki, F. (2011). Improving students' proportional thinking using schema-based instruction. *Learning and Instruction*, *21*(6), 731–745. https://doi.org/10.1016/j.learninstruc.2011.04.002
Jitendra, A. K., Alghamdi, A., Edmunds, R., McKevett, N. M., Mouanoutoua, J., & Roesslein, R. (2021). The Effects of Tier 2 Mathematics Interventions for Students With Mathematics Difficulties: A Meta-Analysis. *Exceptional Children*, *87*(3), 307–325.

*Jordan, N. C., Dyson, N., & Glutting, J. (2011, September). *Developing number sense in kindergartners at risk for learning disabilities in mathematics* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*(3), 647–660. https://doi.org/10.1037/a0029018
*Kanive, R., Nelson, P. M., Burns, M. K., & Ysseldyke, J. (2014). Comparison of the effects of computer-based practice and conceptual understanding interventions on mathematics fact retention and generalization. *The Journal of Educational Research*, *107*(2), 83–89. https://doi.org/10.1080/00220671.2012.759405
*Kariuki, P., & Gentry, C. (2010, November). *The effects of accelerated math utilization on grade equivalency score at a selected elementary school* [Paper presentation]. Paper presented at the annual conference of the Mid-South Educational Research Association, Mobile, AL. https://files.eric.ed.gov/fulltext/ED513432.pdf
*Kariuki, P. N., & Humphrey, S. G. (2006, November). *The effects of drama on the performance of at-risk elementary math students* [Paper presentation]. Paper presented at the annual meeting of the Mid-South Educational Research Association, Birmingham, AL.
*Kebritchi, M. (2008). *Effects of a computer game on mathematics achievement and class motivation: An experimental study* [Doctoral dissertation]. http://etd.fcla.edu/CF/CFE0002066/Kebritchi_Mansureh_200805_PhD.pdf
*Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education*, *55*(2), 427–443. https://doi.org/10.1016/j.compedu.2010.02.007
*Kellman, P. J., Massey, C., Roth, Z., Burke, T., Zucker, J., Saw, A., Aguerob, K. E., & Wise, J. A. (2008). Perceptual learning and the technology of expertise: Studies in fraction learning and algebra. *Pragmatics & Cognition*, *16*(2), 356–405.
Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. National Academy Press.
*King, C. (2012). *Achievement gap: The impact of increased engagement* [Unpublished doctoral dissertation]. Oklahoma State University, Stillwater, OK.
*Kisker, E. E., Lipka, J., Adams, B. L., Rickard, A., Andrew-Ihrke, D., Yanez, E. E., & Millard, A. (2012). The potential of a culturally based supplemental mathematics curriculum to improve the mathematics performance of Alaska Native and other students. *Journal for Research in Mathematics Education*, *43*(1), 75–113. https://doi.org/10.5951/jresematheduc.43.1.0075
*Klein, A., Starkey, P., Clements, D., Sarama, J., & Iyer, R. (2008). Effects of a pre-kindergarten mathematics intervention: A randomized experiment. *Journal of Research on Educational Effectiveness*, *1*(3), 155–178. https://doi.org/10.1080/19345740802114533
*Konstantopoulos, S., Li, W., Miller, S. R., & van der Ploeg, A. (2016). Effects of interim assessments across the achievement distribution: Evidence from an experiment. *Educational and Psychological Measurement*, *76*(4), 587–608. https://doi.org/10.1177/0013164415606498
*Konstantopoulos, S., Li, W., Miller, S. R., & van der Ploeg, A. (2017). Do interim assessments reduce the race and SES achievement gaps? *The Journal of Educational Research*, *110*(4), 319–330. https://doi.org/10.1080/00220671.2015.1103685
*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*(sup1), 188–208. https://doi.org/10.1080/19345747.2015.1116031
*Konstantopoulos, S. L., Miller, S., van der Ploeg, A., Li, C.-H., & Traynor, A. (2011, March). *The impact of Indiana's system of diagnostic assessments on mathematics achievement* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Konstantopoulos, S. L., Wei, L., Miller, S., & van der Ploeg, A. (2015, March). *Effects of interim assessments on the achievement gap: Evidence from an experiment* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Kosko, K. W., & Ferdig, R. E. (2016). Effects of a tablet-based mathematics application for preschool children. *Journal of Computers in Mathematics and Science Teaching*, *35*(1), 61–79.
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
Kraft, M. A., & Falken, G. (2021). *A blueprint for scaling tutoring across public schools* (EdWorkingPaper: 20-335). Annenberg Institute at Brown University.
*Krawec, J., & Huang, J. (2017). Modifying a research-based problem-solving intervention to improve the problem-solving performance of fifth and sixth graders with and without learning disabilities. *Journal of Learning Disabilities*, *50*(4), 468–480. https://doi.org/10.1177/0022219416645565
*Krawec, J., Huang, J., Montague, M., Kressler, B., & de Alba, A. (2013). The effects of cognitive strategy instruction on knowledge of math problem-solving processes of middle school students with learning disabilities. *Learning Disability Quarterly*, *36*(2), 80–92. https://doi.org/10.1177/0731948712463368
Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial and Special Education*, *24*(2), 97–114. https://doi.org/10.1177/07419325030240020501
Kwong, D., & Davis, J. R. (2015). School Climate for Academic Success: A Multilevel Analysis of School Climate and Student Outcomes. *Journal of Research in Education, 25*(2), 68–81.
*Lang, L., & LaVenia, M. (2015). *MFAS 2012-13 year-long study results.* Florida Center for Research in Science, Technology, Engineering, and Mathematics. http://www.cpalms.org/resource/mfas.aspx
*Lang, L., Schoen, R. C., LaVenia, M., & Oberlin, M. (2014, March). *Mathematics formative assessment system—Common core state standards: A randomized field trial in kindergarten and first grade* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
Langdon, D., McKittrick, G., Beede, D., Kahn, B., & Doms, M. (2011). *STEM: Good jobs now and for the future* (ESA Issue Brief 03-11). U.S. Department of Commerce. https://www.purdue.edu/hhs/hdfs/fii/wp-content/uploads/2015/07/s_iafis04c01.pdf
*Lavenia, M. (2016). *Mathematics formative assessment system: Testing the theory of action based on the results of a randomized field trial* [Unpublished doctoral dissertation]. Florida State University, Tallahassee, FL.
Li, X., Dusseldorp, E., Su, X., & Meulman, J. J. (2020). Multiple moderator meta-analysis using the R-package meta-CART. *Behavior Research Methods*, *52*(6), 2657–2673. https://doi.org/10.3758/s13428-020-01360-0
Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review, 22*, 213–243. https://doi.org/10.1007/s10648-010-9125-8
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE.
*Llorente, C., Pasnik, S., Moorthy, S., Hupert, N., Rosenfeld, D., & Gerard, S. (2015). March). *Preschool teachers can use a PBS KIDS transmedia curriculum supplement to support young children's mathematics learning: Results of a randomized controlled trial* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
Lloyd, J. D. (2013, January 1). *Effects of mathematics interventions on elementary students' math skills: A meta-analysis.* University of California, Riverside. http://escholarship.org/uc/item/03m529ts
*Loehr, A. M., & Rittle-Johnson, B. (2017). Putting the "th" in tenths: Providing place-value labels helps reveal the structure of our base-10 numeral system. *Journal of Cognition and Development*, *18*(2), 226–245. https://doi.org/10.1080/15248372.2016.1243118

*Lonigan, C. J., Phillips, B. M., Clancy, J. L., Landry, S. H., Swank, P. R., Assel, M., Taylor, H. B., Klein, A., Starkey, P., Domitrovich, C. E., Eisenberg, N., de Villiers, J., de Villiers, P., & Barnes, M. (2015). Impacts of a comprehensive school readiness curriculum for preschool children at risk for educational difficulties. *Child Development*, *86*(6), 1773–1793. https://doi.org/10.1111/cdev.12460

*Lore, M. D., Wang, A. H., & Buckley, M. T. (2016). Effectiveness of a parent-child home numeracy intervention on urban Catholic school first grade students. *Journal of Catholic Education*, *19*(3), 142–165. https://doi.org/10.15365/joce.1903082016

*Ludwig, M., & Song, M. (2016). *Evaluation of professional development in the use of arts-integrated activities with mathematics content: Findings from the evaluation of the Wolf Trap arts in education model development and dissemination grant*. American Institutes for Research.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293. https://doi.org/10.3102/0162373719849044

*Lynch, K., & Kim, J. S. (2017). Effects of a summer mathematics intervention for low-income children: A randomized experiment. *Educational Evaluation and Policy Analysis*, *39*(1), 31–53. https://doi.org/10.3102/0162373716662339

*Martin, T. B., Brasiel, S. J., Turner, H., & Wise, J. C. (2012). *Effects of the Connected Mathematics Project 2 (CMP2) on the mathematics achievement of Grade 6 students in the mid-Atlantic region* (NCEE 2012-4017). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Mathur, M. B., & VanderWeele, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*, *38*(8), 1336–1342. https://doi.org/10.1002/sim.8057

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119. https://doi.org/10.1111/rssc.12440

*Mattera, S. K., & Morris, P. A. (2017). *Counting on early math skills: Preliminary kindergarten impacts of the Making Pre-K Count and High 5s programs.* MDRC. https://www.mdrc.org/sites/default/files/Making_Pre-K_Count_Brief.pdf

*McClung, L. W. (1998). *A study on the use of manipulatives and their effect on student achievement in a high school Algebra I class* [Unpublished master's thesis]. Salem-Teikyo University, Salem, WV.

*McCoach, D. B., Gubbins, E. J., Foreman, J., Rambo, K. E., & Rubenstein, L. D. (2013, March). *Evaluating the efficacy of using predifferentiated and enriched mathematics curricula for Grade 3 students* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*McCoach, D. B., Gubbins, E. J., Foreman, J., Rubenstein, L. D., & Rambo-Hernandez, K. E. (2014). Evaluating the efficacy of using predifferentiated and enriched mathematics curricula for Grade 3 students: A multisite cluster-randomized trial. *Gifted Child Quarterly*, *58*(4), 272–286. https://doi.org/10.1177/0016986214547631

*McNeil, N. M., Fyfe, E. R., & Dunwiddie, A. E. (2015). Arithmetic practice can be modified to promote understanding of mathematical equivalence. *Journal of Educational Psychology*, *107*(2), 423–436. https://doi.org/10.1037/a0037687

*McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing $4 = 2 + 2$: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development*, *82*(5), 1620–1633.

*Miller, G. I., Jaciw, A., Hoshiko, B., & Wei, X. (2007). *Comparative effectiveness of TI-84 graphing calculators on Algebra I and geometry outcomes: A report of randomized experiments in the East Side Union High School District and San Diego Unified School District*. Empirical Education Inc. https://www.empiricaleducation.com/pdfs/TIfr.pdf

*Montague, M., Krawec, J., Enders, C., & Dietz, S. (2014). The effects of cognitive strategy instruction on math problem solving of middle-school students of varying ability. *Journal of Educational Psychology*, *106*(2), 469–481. https://doi.org/10.1037/a0035176

*Morris, P. A., Mattera, S. K., & Maier, M. F. (2016). *Making pre-K count: Improving math instruction in New York City*. MDRC. https://www.mdrc.org/sites/default/files/Making_Pre-K_Count_FR.pdf
National Center for Education Statistics. (2020). *NAEP: What does the NAEP civics assessment measure?* U.S. Department of Education, National Assessment of Educational Progress. https://nces.ed.gov/nationsreportcard/civics/whatmeasure.aspx
National Center for Education Statistics. (2021). *National Assessment of Educational Progress (NAEP), various years, 1990–2019*. Author. https://www.nationsreportcard.gov/mathematics/nation/achievement/?grade=12
National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Author. https://www.edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf
National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Author.
National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Author.
National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Author.
National Council of Teachers of Mathematics. (2007). *Curriculum focal points for prekindergarten through Grade 8 mathematics: A quest for coherence*. Author.
National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Authors.
National Science and Technology Council. (2013). *Federal science, technology, engineering and mathematics (STEM) education: 5-Year strategic plan*. Author. https://www.whitehouse.gov/sites/default/files/microsites/ostp/stem_stratplan_2013.pdf
National Science and Technology Council. (2018). *Charting a course for success: America's strategy for STEM education*. Author. https://www.whitehouse.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf
*Nelson, P. M., Burns, M. K., Kanive, R., & Ysseldyke, J. E. (2013). Comparison of a math fact rehearsal and a mnemonic strategy approach for improving math fact fluency. *Journal of School Psychology*, *51*(6), 659–667. https://doi.org/10.1016/j.jsp.2013.08.003
Nelson, G., & McMaster, K. L. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. *Journal of Educational Psychology, 111*(6), 1001–1022. https://doi.org/10.1037/edu0000334
*Newman, D. F., Pamela, 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, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence* (EdWorkingPaper: 20-267). Annenberg Institute at Brown University. https://doi.org/10.26300/eh0c-pc52
*Nishida, T. K. (2008). The use of manipulatives to support children's acquisition of abstract math concepts. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *69*(1-B), 718.
Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237
*Omniewski, R. A., & Habursky, B. (1999). Does arts infusion make a difference? The effect of an arts infusion approach on mathematics achievement. *Contributions to Music Education*, *25*(2), 38–50.
*Orabuchi, I. I. (1992). *Effects of using interactive CAI on primary grade students' higher-order thinking skills: Inferences, generalizations, and math problem solving* [Unpublished doctoral dissertation]. Texas Woman's University.
Organisation for Economic Co-operation and Development. (2021). *Mathematics performance (PISA) (indicator)*. https://doi.org/10.1787/04711c74-en

*Owen, R. L., & Fuchs, L. S. (2002). Mathematical problem-solving strategy instruction for third-grade students with learning disabilities. *Remedial and Special Education*, *23*(5), 268–278. https://doi.org/10.1177/07419325020230050201

*Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127–144. https://doi.org/10.3102/0162373713507480

*Pane, J. F., McCaffrey, D. F., Ikemoto, G. S., Steele, J. L., & Slaughter, M. E. (2009, March). *Results from a randomized efficacy trial of Cognitive Tutor Geometry* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Pane, J. F., McCaffrey, D. F., Slaughter, M. E., Steele, J. L., & Ikemoto, G. S. (2010). An experiment to evaluate the efficacy of Cognitive Tutor Geometry. *Journal of Research on Educational Effectiveness*, *3*(3), 254–281. https://doi.org/10.1080/19345741003681189

*Pape, S. J., Irving, K. E., Owens, D. T., Boscardin, C. K., Sanalan, V. A., Abrahamson, A. L., Kaya, S., Shin, H. S., & Silver, D. (2012). Classroom connectivity in Algebra I classrooms: Results of a randomized control trial. *Effective Education*, *4*(2), 169–189. https://doi.org/10.1080/19415532.2013.841059

*Pare-Blagoev , Booth, J., Elliot, A., & Koedinger, K. (2013, September). *Using worked example assignments in classroom instruction* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

*Parlapanides, T. (2009). *Effects of a technology treatment on student scores on the standardized Grade 8 Proficiency Assessment (GEPA) in New Jersey* [Doctoral dissertation]. https://scholarship.shu.edu/dissertations/387

*Parr, B., Edwards, M. C., & Leising, J. G. (2009). Selected effects of a curriculum integration intervention on the mathematics performance of secondary students enrolled in an agricultural power and technology course: An experimental study. *Journal of Agricultural Education*, *50*(1), 57–69. https://doi.org/10.5032/jae.2009.01057

*Parr, B. A., Edwards, M. C., & Leising, J. G. (2006). Effects of a math-enhanced curriculum and instructional approach on the mathematics achievement of agricultural power and technology students: An experimental study. *Journal of Agricultural Education*, *47*(3), 81–93. https://doi.org/10.5032/jae.2006.03081

*Pasnak, R., Hansbarger, A., Dodson, S. L., Hart, J. B., & Blaha, J. (1996). Differential results of instruction at the preoperational/concrete operational transition. *Psychology in the Schools*, *33*(1), 70–83. https://doi.org/10.1002/(SICI)1520-6807(199601)33:1<70::AID-PITS9>3.0.CO;2-#

*Pasnik, S., & Llorente, C. (2013). *Preschool teachers can use a PBS KIDS transmedia curriculum supplement to support young children's mathematics learning: Results of a randomized controlled* trial. A report to the CPB-PBS Ready to Learn Initiative. Education Development Center and SRI International.

*Pasnik, S., Moorthy, S., Llorente, C., Hupert, N., Dominguez, X., & Silander, M. (2015). *Supporting parent-child experiences with PEG + CAT Early Math Concepts. A report to the CPB-PBS Ready to Learn Initiative.* Education Development Center and SRI International. http://cct.edc.org/sites/cct.edc.org/files/ms-resources/edc-sri-rtl-peg-math-study-report-2015.pdf

*Pearson, D. (2004). Working the math: Professional development in the NRCCTE Math-in-CTE project. *Techniques: Connecting Education and Careers*, *79*(6), 22–23.

*Phelan, J. C., Choi, K., Niemi, D., Vendlinski, T. P., Baker, E. L., & Herman, J. (2012). The Effects of POWERSOURCE assessments on middle-school students' math performance. *Assessment in Education: Principles, Policy, and Practice*, *19*(2), 211–230. https://doi.org/10.1080/0969594X.2010.532769

*Phelan, J., Choi, K., Vendlinski, T., Baker, E. L., & Herman, J. L. (2009). *The effects of POWERSOURCE© intervention on student understanding of basic mathematical principles (CRESST Report 763).* University of California, National Center for Research on Evaluation, Standards, and Student Testing.

*Phelan, J., Choi, K., Vendlinski, T., Baker, E., & Herman, J. (2011). Differential improvement in student understanding of mathematical principals following formative assessment intervention. *The Journal of Educational Research*, *104*(5), 330–339. https://doi.org/10.1080/00220671.2010.484030

Pigott, T. D. (2001). Missing predictors in models of effect size. *Evaluation & the Health Professions*, *24*(3), 277–307. https://doi.org/10.1177/01632780122034920

Pigott, T. D. (2012). Missing data in meta-analysis: Strategies and approaches. In *Advances in meta-analysis* (pp. 79–107). Springer. https://doi.org/10.1007/978-1-4614-2278-5_7

*Ploger, D., & Hecht, S. (2009). Enhancing children's conceptual understanding of mathematics through chartworld software. *Journal of Research in Childhood Education*, *23*(3), 267–277. https://doi.org/10.1080/02568540909594660

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

*Powell, S. R., Driver, M. K., & Julian, T. E. (2015). The effect of tutoring with nonstandard equations for students with mathematics difficulty. *Journal of Learning Disabilities*, *48*(5), 523–534. https://doi.org/10.1037/a0028389

*Powers, S., & Price-Johnson, C. (2007). *Evaluation of the Waterford Early Math & Science Program for kindergarten: First-year implementation in five urban low-income schools*. Creative Research Associates, Inc.

*Presser, A. L., Clements, M., Ginsburg, H., & Ertle, B. (2015). Big math for little kids: The effectiveness of a preschool and kindergarten mathematics curriculum. *Early Education and Development*, *26*(3), 399–426. https://doi.org/10.1080/10409289.2015.994451

President's Council of Advisors on Science and Technology. Report to the President: Engage to Excel: Producing One Million New College Graduates with Degrees in Science, Technology, Engineering and Mathematics. (2012). https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf

*Purpura, D. J., Napoli, A. R., Wehrspann, E. A., & Gold, Z. S. (2017). Causal connections between mathematical language and mathematical knowledge: A dialogic reading intervention. *Journal of Research on Educational Effectiveness*, *10*(1), 116–137. https://doi.org/10.1080/19345747.2016.1204639

Pustejovsky, J.E., & Tipton, E. (2021). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*. https://doi.org/10.1007/s11121-021-01246-3

Pustejovsky, J. (2017). *Pooling clubSandwich results across multiple imputations* [Blog post]. https://www.jepusto.com/mi-with-clubsandwich/

Pustejovsky, J. (2018). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.3.2). https://CRAN.R-project.org/package=clubSandwich

Quartagno, M., Grund, S., & Carpenter, J. (2019). jomo: A flexible package for two-level joint modelling multiple imputation. *The R Journal*, *11*(2), 205–228. https://doi.org/10.32614/RJ-2019-028

*Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2011). *Classroom assessment for student learning: Impact on elementary school mathematics in the central region* (NCEE 2011-4005). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

*Rast, J. D. (2005). *A comparison of learning subjective and traditional probability in middle grades* [Doctoral dissertation]. https://scholarworks.gsu.edu/msit_diss/4/

*Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561–574. https://doi.org/10.1037/0022-0663.99.3.561

*Rivera, F. (2015). *Effect of instruction and textbook adoption procedures on kindergarten students' learning of the concept of rectangle* [Unpublished doctoral dissertation]. The University of Texas-Pan American, Edinburg, TX.

*Rolfhus, E., Gersten, R., Clarke, B., Decker, L., Wilkins, C., & Dimino, J. (2012). *An evaluation of Number Rockets: A Tier 2 intervention for Grade 1 students at risk for difficulties in mathematics* (NCEE 2012-4007). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

*Rolfhus, E., Gersten, R., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2013). March). *From efficacy trial to large scale effectiveness trial: A Tier 2 mathematics intervention for first graders with difficulties in mathematics* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

*Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*(4), 833–878. https://doi.org/10.3102/0002831210367426

*Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J., Richland, L., Tran, N., Schneider, S., Duran, L., & Martinez, M. 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

*Rutherford, T., Kibrick, M., Burchinal, M., Richland, L., Conley, A., Osborne, K., Schneider, S., Duran, L., Coulson, A., Antenore, F., Daniels, A., &, Martinez, M. E. (2010, May). *Spatial temporal mathematics at scale: An innovative and fully developed paradigm to boost math achievement among all learners* [Paper presentation]. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Sahni, S. D., Polanin, J. R., Zhang, Q., Michaelson, L. E., Caverly, S., Polese, M. L., & Yang, J. (2021). *A What Works Clearinghouse rapid evidence review of distance learning programs* (WWC 2021-005REV). What Works Clearinghouse.

*Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2010). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness*, *4*(1), 1–24. https://doi.org/10.1080/19345747.2010.498562

*Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *1*(2), 89–119. https://doi.org/10.1080/19345740801941332

*Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2012). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *5*(2), 105–135. https://doi.org/10.1080/19345747.2011.627980

Schneider, M. (2021, June 29). To build a STEM workforce, we must invest in education science. But a bill congress is considering doesn't go far enough. *The 74*. https://www.the74million.org/article/schneider-to-build-a-stem-workforce-we-must-invest-in-education-science-but-a-bill-congress-is-considering-doesnt-go-far-enough/

*Schenke, K., Rutherford, T., & Farkas, G. (2014). Alignment of game design features and state mathematics standards: Do results reflect intentions? *Computers & Education*, *76*, 215–224. https://doi.org/10.1016/j.compedu.2014.03.019

*Shoemaker, T. L. (2013). *Effect of computer-aided instruction on attitude and achievement of fifth grade math students* [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 3595355). http://search.proquest.com/docview/1443878124

*Shults, P. A. (2000). *Teaching first grade computation: A comparison of traditional instruction and computer enhanced instruction* [Unpublished master's thesis]. Johnson Bible College, Kimberlin Heights, TN.

*Silander, M., Moorthy, S., Dominguez, X., Hupert, N., Pasnik, S., & Llorente, C. (2016). March). *Using digital media at home to promote young children's mathematics learning: Results of a randomized controlled trial* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Silvious, N. B. (2008). *Effects of Saxon Math program of instruction on the mathematics achievement of students with learning disabilities in Grades 2 through 8* [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 3363773). https://search.proquest.com/docview/304802200

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, *51*(7), 870–901. https://doi.org/10.1002/tea.21139

*Smith, J. G. (2012). *Screen-capture instructional technology: A cognitive tool for blended learning* [Unpublished doctoral dissertation]. Saint Mary's College of California, Moraga, CA.

*Snipes, J., Huang, C.-W., Jaquet, K., & Finkelstein, N. (2015). *The effects of the ElevateMath summer program on math achievement and algebra readiness* (REL 2015–096). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. http://ies.ed.gov/ncee/edlabs

*Snipes, J., Huang, C.-W., Jaquet, K., & Finkelstein, N. (2016, March). *The effects of the Elevate Math summer program on math achievement and algebra readiness* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Sood, S., & Jitendra, A. K. (2013 Jul-Aug). An exploratory study of a number sense program to develop kindergarten students' number proficiency. *Journal of Learning Disabilities*, *46*(4), 328–346.

*Sorensen, C. (2011). Learning with mobile technologies: *The use of out-of-class short message system text messaging to support the classroom learning of high school algebra* [Unpublished doctoral dissertation]. University of South Florida, Tampa, FL.

*Spotnitz, S. H. (2001). *Intrinsic motivation in students with learning disabilities as examined through computer-based instruction in mathematics* [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 3005801).

*Springer, R., Pugalee, D., & Algozzine, B. (2007). Improving mathematics skills of high school students. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *81*(1), 37–44. https://doi.org/10.3200/TCHS.81.1.37-44

*Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, *40*, 41–54. https://doi.org/10.1016/j.cedpsych.2014.05.005

*Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology*, *102*(4), 408–426. https://doi.org/10.1016/j.jecp.2008.11.004

*Star, J. R., Rittle-Johnson, B., Durkin, K., Newton, K., Pollack, C., Lynch, K., & Gogolen, C. (2013). March). *The impact of a comparison curriculum in Algebra I: A randomized experiment* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, *105*(4), 970–987. https://doi.org/10.1037/a0032447

*Stone, J. R., III, Alfeld, C., & Pearson, D. (2008). Rigor and relevance: Enhancing high school students' math skills through career and technical education. *American Educational Research Journal*, *45*(3), 767–795. https://doi.org/10.3102/0002831208317460

*Stone, J. R., III, Alfeld, C., Pearson, D., Lewis, M. V., & Jensen, S. (2005). *Building academic skills in context: Testing the value of enhanced math learning in CTE Pilot Study*. University of Minnesota, National Research Center for Career and Technical Education. https://files.eric.ed.gov/fulltext/ED497344.pdf

*Stone, J. R., III, Alfeld, C., Pearson, D., Lewis, M. V., & Jensen, S. (2006). *Building academic skills in context: Testing the value of enhanced math learning in CTE*. National Dissemination Center for Career and Technical Education. http://files.eric.ed.gov/fulltext/ED493604.pdf

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

*Swanson, H. L. (2014). Does cognitive strategy training on word problems compensate for working memory capacity in children with math difficulties? *Journal of Educational Psychology*, *106*(3), 831–848. https://doi.org/10.1037/a0035838
*Swanson, H. L., Lussier, C., & Orosco, M. (2013). Effects of cognitive strategy interventions and cognitive moderators on word problem solving in children at risk for problem solving difficulties. *Learning Disabilities Research & Practice*, *28*(4), 170–183. https://doi.org/10.1111/ldrp.12019
*Swanson, H. L., Moran, A., Lussier, C., & Fung, W. (2014). The effect of explicit and direct generative strategy training and working memory on word problem-solving accuracy in children at risk for math difficulties. *Learning Disability Quarterly*, *37*(2), 111–123. https://doi.org/10.1177/0731948713507264
*Swanson, H. L., Orosco, M. J., & Lussier, C. (2013, March). *Does cognitive strategy training on word problems compensate for working memory capacity in children with math difficulties?* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.
*Swanson, H. L., Orosco, M. J., & Lussier, C. M. (2014). The effects of mathematics strategy instruction for children with serious problem-solving difficulties. *Exceptional Children*, *80*(2), 149–168. https://doi.org/10.1177/001440291408000202
*Taub, G. E., McGrew, K. S., & Keith, T. Z. (2015). Effects of improvements in interval timing on the mathematics achievement of elementary school students. *Journal of Research in Childhood Education*, *29*(3), 352–366. https://doi.org/10.1080/02568543.2015.1040563
Taylor, J. A., Kowalski, S. M., Polanin, J. R., Askinas, K., Stuhlsatz, M. A. M., Wilson, C. D., Tipton, E., & Wilson, S. J. (2018). Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open*, *4*(3), 233285841879199–233285841879119. https://doi.org/10.1177/2332858418791991
*Tieso, C. (2005). The effects of grouping practices and curricular adjustments on achievement. *Journal for the Education of the Gifted*, *29*(1), 60–89. https://doi.org/10.1177/016235320502900104
*Tieso, C. L. (2002). *The effects of grouping and curricular practices on intermediate students' math achievement* (RM02154). National Research Center on the Gifted and Talented.
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
Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099
Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161–179. https://doi.org/10.1002/jrsm.1338
*Torlakovic, E. (2011). *Academy of MATH® efficacy study: Randomized control study.* EPS School Specialty Literacy and Intervention.
*U.S. Department of Education. (2010a). *What Works Clearinghouse intervention report: Accelerated Math*™. https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_accelmath_091410.pdf
*U.S. Department of Education. (2010b). *What Works Clearinghouse intervention report: PLATO® Achieve Now.* https://files.eric.ed.gov/fulltext/ED508582.pdf
*U.S. Department of Education. (2010c). *What Works Clearinghouse quick review of the report Middle school mathematics professional development impact study: Findings after the first year of implementation.* https://files.eric.ed.gov/fulltext/ED511482.pdf
*U.S. Department of Education. (2012). *What Works Clearinghouse review of the report A multi-site cluster randomized trial of the effects of CompassLearning Odyssey® Math on the math achievement of selected Grade 4 students in the Mid-Atlantic region.* https://ies.ed.gov/ncee/wwc/Docs/SingleStudyReviews/wwc_odysseymath_053012.pdf
*U.S. Department of Education. (2013a). *What Works Clearinghouse intervention report: DreamBox Learning.* https://files.eric.ed.gov/fulltext/ED544762.pdf

*U.S. Department of Education. (2013b). What Works Clearinghouse review of the report Improving at-risk learners' understanding of fractions. https://files.eric.ed.gov/fulltext/ED544204.pdf

*U.S. Department of Education. (2014a). *What Works Clearinghouse review of the report Benefits of practicing 4 = 2 + 2: Nontraditional problem formats facilitate children's understanding of mathematical equivalence.* https://files.eric.ed.gov/fulltext/ED544791.pdf

*U.S. Department of Education. (2014b). *What Works Clearinghouse review of the report The effects of cognitive strategy instruction on math problem solving of middle school students of varying ability.* https://files.eric.ed.gov/fulltext/ED544795.pdf

*U.S. Department of Education. (2014c). *What Works Clearinghouse review of the report Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies.* https://files.eric.ed.gov/fulltext/ED545228.pdf

*U.S. Department of Education. (2015). *What Works Clearinghouse review of the report The impact of Indiana's system of interim assessments on mathematics and reading achievement.* https://files.eric.ed.gov/fulltext/ED560749.pdf

*U.S. Department of Education. (2016). *What Works Clearinghouse intervention report: Cognitive Tutor®.* https://files.eric.ed.gov/fulltext/ED566735.pdf

*Uttal, D. H., Amaya, M., del Rosario Maita, M., Hand, L. L., Cohen, C. A., O'Doherty, K., & DeLoache, J. S. (2013). It works both ways: Transfer difficulties between manipulatives and written subtraction solutions. *Child Development Research*, *2013*, 1–13. https://doi.org/10.1155/2013/216367

*Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., Ruiz de Castilla, V., & Sullivan, K. (2015, March). Preliminary findings from a multi-year scale-up effectiveness trial of *Everyday Mathematics* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., Ruiz de Castilla, V., Sullivan, K., & Rodriguez, D. (2015). *Findings from a multi-year scale-up effectiveness trial of Everyday Mathematics* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

van Lissa, C. J. (2017). *MetaForest: Exploring heterogeneity in meta-analysis using random forests.* https://doi.org/10.31234/osf.io/myg6s

van Lissa, C. J. (2020). Small sample meta-analyses: Exploring heterogeneity using MetaForest. In R. Van De Schoot & M. Miočević (Eds.), *Small sample size solutions (open access): A guide for applied researchers and practitioners.* CRC Press. https://www.crcpress.com/Small-Sample-Size-Solutions-Open-Access-A-Guide-for-Applied-Researchers/Schoot-Miocevic/p/book/9780367222222

*Van Voorhis, F. L. (2011). Adding families to the homework equation: A longitudinal study of mathematics achievement. *Education and Urban Society*, *43*(3), 313–338. https://doi.org/10.1177/0013124510380236

*VanDerHeyden, A., & Codding, R. S. (2015). Practical effects of classwide mathematics intervention. *School Psychology Review*, *44*(2), 169–190. https://doi.org/10.17105/spr-13-0087.1

*VanDerHeyden, A., McLaughlin, T., Algina, J., & Snyder, P. (2012). Randomized evaluation of a supplemental gradewide mathematics intervention. *American Educational Research Journal*, *49*(6), 1251–1284. https://doi.org/10.3102/0002831212462736

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989x.10.4.428

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

*Wang, H., & Woodworth, K. (2011). *Evaluation of Rocketship Education's use of DreamBox Learning's online mathematics program*. SRI International. https://www.dreambox.com/wp-content/uploads/downloads/pdf/DreamBox_Results_from_SRI_Rocketship_Evaluation.pdf

*Wang, H., & Woodworth, K. (2011, September). *A randomized controlled trial of two online mathematics curricula* [Paper presentation]. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

*Washington, L. (2015). *The relationship of students' perceptions of classroom environment on mathematically promising English language learners on math achievement* [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 3663075).

*Watt, S. J. (2013). *Teaching algebra-based concepts to students with learning disabilities: The effects of preteaching using a gradual instructional sequence* [Doctoral dissertation]. https://ir.uiowa.edu/etd/2658

*Watt, S. J., & Therrien, W. J. (2016). Examining a preteaching framework to improve fraction computation outcomes among struggling learners. *Preventing School Failure: Alternative Education for Children and Youth*, *60*(4), 311–319. https://doi.org/10.1080/1045988X.2016.1147011

*Weaver, C. L. (1991). *Young children learn geometric and spatial concepts using Logo with a screen turtle and a floor turtle* [Doctoral dissertation]. Reproduced from microfilm master (Order Number 9135151).

What Works Clearinghouse. (2020). *Procedures and standards handbooks (version 4.1)*. https://ies.ed.gov/ncee/wwc/Handbooks

Wheeler, J. L., & Regian, J. W. (1999). The use of a cognitive tutoring system in the improvement of the abstract reasoning component of word problem solving. *Computers in Human Behavior*, *15*(2), 243–254. https://doi.org/10.1016/S0747-5632(99)00021-7

White House. (2012). *Report to the president, engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Executive Office of the President, President's Council of Advisors on Science and Technology. https://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf

*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: Final report* (NCEE 2009-4068). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

*Wilcox, D. M. (1997). *The use of animation with instruction and feedback in fractions software for children* [Doctoral dissertation]. ProQuest Dissertations and Theses database (UMI No. 9803666).

*Wilder, S., & Berry, L. (2016). Emporium Model: The key to content retention in secondary math courses. *Journal of Educators Online*, *13*(2), 53–75.

*Winburg, K., Chamberlin, B., Valdez, A., Trujillo, K., & Stanford, T. B. (2016). Impact of math snacks games on students' conceptual understanding. *Journal of Computers in Mathematics and Science Teaching*, *35*(2), 173–193.

Wolf, R. (2021). Average differences in effect sizes by outcome measure type. What Works Clearinghouse. https://eric.ed.gov/?id=ED610568

*Yang, J. (2013). *The influential factors of math achievement in mathematically promising English language learners* [Unpublished doctoral dissertation]. St. John's University, New York, NY.

*Young, R. B., Hodge, A., Edwards, M. C., & Leising, J. G. (2012). Learning mathematics in high school courses beyond mathematics: Combating the need for post-secondary remediation in mathematics. *Career and Technical Education Research*, *37*(1), 21–33. https://doi.org/10.5328/cter37.1.21

*Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review*, *36*(3), 453–467. https://doi.org/10.1080/02796015.2007.12087933

Zollman, A. (2012). Learning for STEM literacy: STEM literacy for learning. *School Science and Mathematics*, *112*(1), 12–19. https://doi.org/10.1111/j.1949-8594.2012.00101.x