Abstract Title Page

Title: Heterogeneity in Mathematics Intervention Effects: Evidence from A Meta-Analysis of 191 Randomized Experiments

Authors and Affiliations:
Martyna Citkowicz, American Institutes for Research
Jim Lindsay, American Institutes for Research
David Miller, American Institutes for Research
Ryan Williams, (presenter) American Institutes for Research

Abstract Body

Background/Context:
Improving the education of America's youth in the disciplines of science, technology, engineering, and mathematics (STEM) is a well-documented, widely-endorsed federal policy priority (U.S. Department of Education, 2017). 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. Yet, too few young Americans attain postsecondary degrees in STEM fields to meet this demand (Change the Equation, 2015; Langdon, McKittrick, Beede, Kahn, & Doms, 2011).

To address this economic reality, many public and private initiatives have focused on improving prekindergarten (PK) through Grade 12 mathematics education, given that the mathematics content learned before college is foundational for later learning and college success across STEM subjects (Achieve, 2014; Calcagno & Long, 2008). Recent reforms include developing more rigorous mathematics curricula, assessments, and standards like the Common Core mathematics standards released in 2010 (Porter, McMaken, Hwang, & Yang, 2010).

However, because of the magnitude of this problem, no single initiative is likely to resolve it. A sustained, multipronged approach is likely needed, involving changes to education policy, curricula, and the way mathematics content is taught. Although decades of educational research have studied PK–12 mathematics interventions, the field lacks a comprehensive understanding of which interventions work, for whom, and under what conditions.

Purpose:
Our meta-analytic project aimed to examine the heterogeneity in mathematics intervention effects by synthesizing 25 years of randomized experiments of interventions designed to improve mathematics achievement. We seek to answer two broad questions:

1. How heterogeneous are mathematics intervention effects?
2. What factors, such as participant, intervention, and outcome characteristics, explain or contribute to the heterogeneity of intervention effects?

Approach:
To find relevant published and unpublished studies, we searched electronic databases (e.g., ERIC, Google Scholar, PsycINFO, the WWC intervention report database, and the WWC registry of randomized control trials) and websites of research organizations (e.g., Comprehensive and Content Centers, research organizations such as Mathematica Policy Research, MDRC, RAND, National Center on Education Evaluation and Regional Assistance) for experimental studies on PK–12 mathematics interventions.

We included studies meeting the following criteria:

1. Included at least one specific intervention/strategy/program designed to improve the teaching or learning of mathematics;
2. Conducted a randomized control group trial;
3. Included a sample of students in Grades PK–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. Provided sufficient information to calculate an effect size estimate and variance;
6. Written in English; and
7. Published in 1991 or later.

Using the criteria above, we conducted three stage of screening: 1) abstract and title screening, 2) full-text screening, and 3) methods screening. Studies that made it through all three stages of screening were coded for information related to the study (e.g., publication type, year of publication), methods (i.e., research design), samples (e.g., student demographic characteristics, grade levels), interventions (e.g., intervention type, features, delivery), outcomes (e.g., measure type, content domain), settings (e.g., geographic locale, urbanicity), and effect sizes (i.e., summary statistics for the impact estimate).

We conducted a series of robust random-effects meta-regression analyses to quantify effect heterogeneity and study the driving forces of that heterogeneity. We conducted analyses in R using the *metafor* package (Viechtbauer, 2010), adjusted for effect size dependencies using robust variance estimation and the *clubSandwich* package (Pustejovsky, 2019), and accounted for missing moderator data using multiple imputation and the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011).

Findings:
From an initial list of 9,384 unique abstracts and titles to be screened, 2,462 studies made it to full-text screening, 796 made it to methods screening, and 283 were eligible randomized experiments that reported sufficient information to calculate effect sizes. Our initial data analyses focused on the subset of studies that compared a mathematics intervention to a "business as usual" comparison condition, resulting in 1107 effect sizes from 191 studies. These studies represented a diverse range of intervention characteristics, sample demographics, and outcome measures, as shown in Table 1.

The overall effect size was moderate, $g = 0.31$, $SE = 0.03$, $p < 0.001$, and heterogeneity was large ($\tau = 0.44$, based on first combining both between-study and within-study heterogeneity parameters). The estimated middle 95% of true underlying effects (i.e., the prediction interval) fell between -0.55 to 1.17. Figure 1 shows the effect size distributions separately by intervention type (curriculum, pedagogical, supplemental time); see also Table 2.

Table 3 extends these analyses to a broader set of moderators: intervention type, intervention training, intervention length, intervention delivery, outcome type, and publication year, while controlling for methodological confounds (e.g., level of random assignment, attrition). These moderators were selected from a model building process in which we separately examined blocks of demographic, intervention, outcome, and setting moderators, controlling for methods moderators. We selected moderators that had $p < .10$ for the composite model in Table 3.

The moderators in the combined model (Table 3) together explained 12% of the total effect size variance (i.e., heterogeneity). As shown, the outcome measure type was one of the largest contributors of heterogeneity. Researcher- and practitioner-generated measures yielded an average effect of $g = 0.46$, which was more than triple the magnitude for standardized achievement outcomes ($g = 0.15$), controlling for other moderators.

Conclusions:
While average math intervention effects are positive across a range of program types, grade levels, and outcome domains, they are especially heterogeneous. Our paper presentation will provide a comprehensive overview of the drivers of effect heterogeneity in mathematics intervention experiments over the past quarter century. It will describe intersections of interventions, intervention components, and outcomes that appear especially promising for future intervention research.

References

Achieve. (2014). *Rising to the challenge*. Washington, DC: Author. Retrieved from http://www.achieve.org/rising-challenge

Calcagno, J. C., & Long, B. T. (2008). The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance (No. w14194). National Bureau of Economic Research.

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

Langdon, D., McKittrick, G., Khan, B., & Doms, M. (2011). *STEM: Good jobs now and for the future*. U.S. Department of Commerce, Economics and Statistics Administration. Retrieved from http://www.esa.doc.gov/sites/default/files/reports/documents/stemfinalyjuly14_1.pdf

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new US intended curriculum. *Educational Researcher*, *40*(3), 103-116.

Pustejovsky, J. (2019). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.3.2). Retrieved from https://CRAN.R-project.org/package=clubSandwich

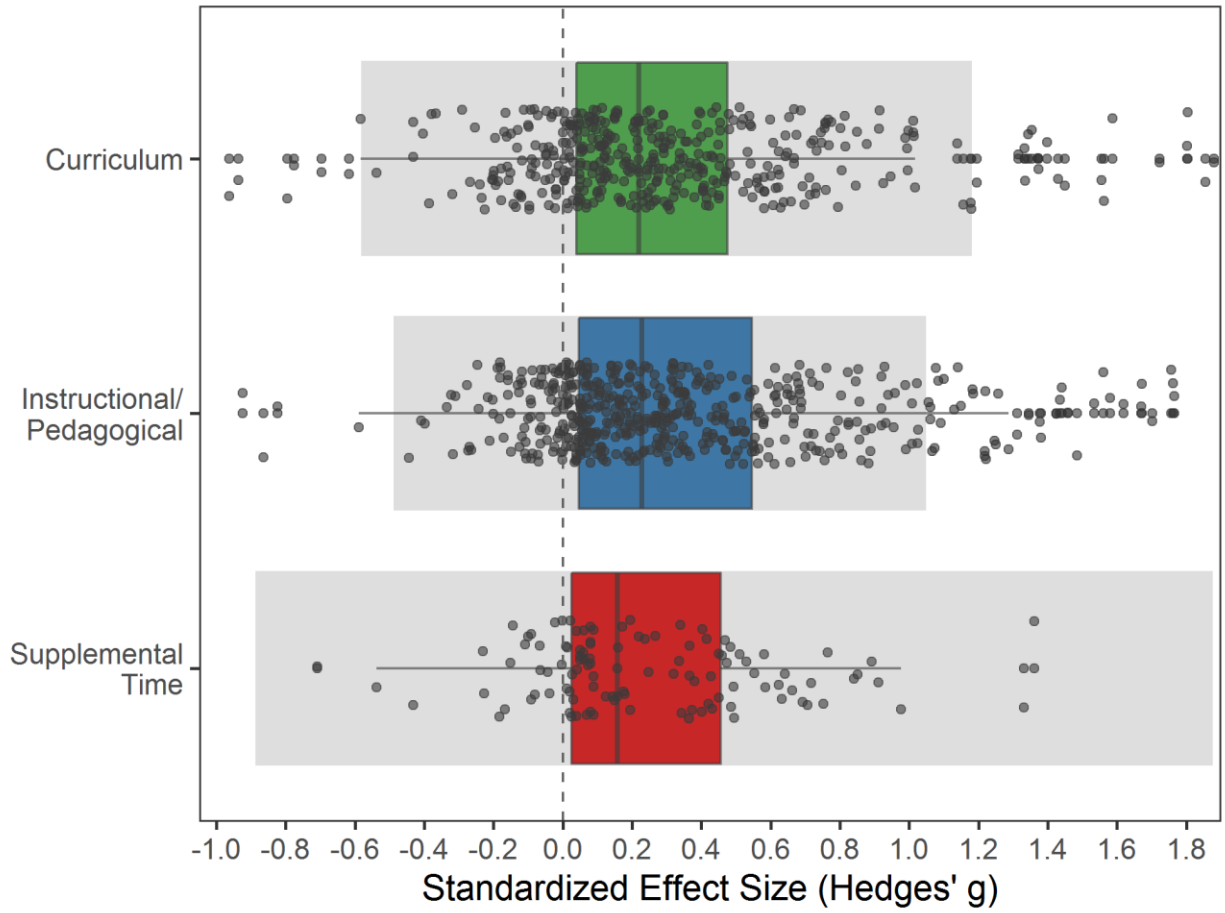
U.S. Department of Education (2017, October 12). Secretary's proposed supplemental priorities and definitions for discretionary grant programs. *Federal Register*, *82*(196), 47484-47493.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Appendix

Figure 1. Boxplot of the Effect Size Distribution by Mathematics Intervention Type



Note. These effect sizes come from 191 randomized experiments comparing student mathematics achievement in an intervention condition relative to a "business as usual" comparison condition. The inner rectangles show the first quartile, median, and third quartile of the observed, unweighted effect size distributions. The outer grey rectangles show the 95% prediction intervals, the estimated middle 95% of true underlying effects.

Table 1. Characteristics of 191 Included Experiments (1,107 Effect Sizes)

Characteristic	*m*	*k*	Mean	Missing (%)
Intervention Type				
Curriculum	82	438	40%	0%
Pedagogical/Instructional	85	553	50%	0%
Supplemental Time	25	116	10%	0%
Intervention Content Domain				
Number Sense & Arithmetic	91	645	66%	12%
Rational Numbers & Fractions	40	198	20%	12%
Algebra & Prealgebra	57	269	28%	12%
Geometry	42	230	24%	12%
Measurement, Data, & Statistics	39	208	21%	12%
Calculus & Precalculus	1	1	0%	12%
Implementation Fidelity				
High	41	395	71%	50%
Medium	22	114	20%	50%
Low	9	48	9%	50%
Random Assignment Level				
Student	93	539	49%	0%
Teacher/Classroom	67	385	35%	0%
School	33	183	17%	0%
District	0	0	0%	0%
Grade Level				
Prekindergarten	18	76	7%	4%
Elementary School	112	771	73%	4%
Middle School	62	241	23%	4%
High School	27	81	8%	4%
Demographics				
% Male	–	–	52%	30%
% Special Education	–	–	20%	71%
% English Language Learner	–	–	22%	65%
% Economically Disadvantaged	–	–	57%	58%
% White	–	–	40%	39%
% Hispanic	–	–	25%	41%
% Black	–	–	32%	38%
% Asian	–	–	6%	57%

Urbanicity					
	Suburban	51	299	45%	40%
	Urban	82	475	72%	40%
	Rural	39	222	34%	40%
U.S. Geographic Region					
	West	35	200	22%	16%
	Midwest	30	188	20%	16%
	Southwest	40	264	28%	16%
	Northeast	52	361	39%	16%
	Southeast	63	329	35%	16%
Outcome Measure Content Domain					
	Number Sense & Arithmetic	92	579	57%	8%
	Rational Numbers & Fractions	39	189	18%	8%
	Algebra & Prealgebra	61	220	22%	8%
	Geometry	47	168	16%	8%
	Measurement, Data, & Statistics	45	154	15%	8%
	Calculus & Precalculus	0	0	0%	8%
Outcome Measure Type					
	Standardized Achievement Test	107	477	43%	0%
	Researcher-Developed Measure	121	628	57%	0%
	Course Credits/Enrollment/Retention	2	2	0.3%	0%

Note. 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 = average percentage for non-missing values (weighted by number of effect sizes), Missing (%) = percentage of effect sizes that have missing values for that characteristic.

Table 2. Random-Effects Meta-Analyses Conducted Separately by Intervention Type

Intervention Type	*m*	*k*	*g*	*SE*	*p*	*τ*	95% Prediction Interval
Curriculum	82	438	0.30	0.04	<.001	0.45	[-0.58, 1.18]
Pedagogical/Instructional	85	553	0.28	0.04	<.001	0.39	[-0.49, 1.05]
Supplemental Time	25	116	0.49	0.15	.003	0.70	[-0.89, 1.88]

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.

m = number of studies, *k* = number of effect sizes, *g* = average effect size, *SE* = standard error of the average effect sizes, *p* = significance level for the mean being different from 0, τ = estimated standard deviation of the true underlying effect sizes, 95% prediction interval = estimated middle 95% of the true underlying effect sizes.

Table 3. Conditional Means from Mixed-Effects Meta-Regression Model

Moderator	*b*	*SE*	*p*
Intervention Type			.008
Curriculum	0.326	0.047	
Pedagogical/instructional	0.249	0.041	
Supplemental	0.673	0.140	
Intervention Training			.273
None or not reported	0.331	0.074	
One-time training	0.248	0.057	
Infrequent ongoing training	0.379	0.062	
Frequent ongoing training	0.393	0.084	
Intervention Length			
Number of weeks*	0.372	0.047	.170
Intervention Delivery			.049
Teacher	0.364	0.044	
Technology	0.103	0.103	
Interventionist	0.391	0.064	
Outcome Type			.002
Not a standardized measure	0.459	0.055	
Standardized achievement measure	0.145	0.055	
Publication Year*	-0.010	0.007	.153

Note. These effect sizes (*g*) represent the predicted means from a multivariable, mixed-effects meta-regression model that simultaneously controlled for all listed moderators (e.g., average effect size for curriculum interventions when the other moderators were fixed at their means across all intervention types). The standard errors were adjusted for effect size dependencies using robust variance estimation. Missing moderator values were handled using multiple imputation; 80 imputed datasets were generated, analyzed separately, and then pooled, using both the within-imputation and between-imputation variance to compute standard errors.

*Indicates regression coefficient rather than conditional mean.