USC
Dornsife
Dana and David Dornsife
College of Letters, Arts
and Sciences
*Center for Applied Research
in Education*

USC
Rossier
EdPolicy Hub

Eureka Math Squared Efficacy Study

December 20, 2024

Daniel Silver, PhD
Morgan S. Polikoff, PhD

Executive summary

This report presents results of a 2024 study that investigated the impact of Eureka Math Squared curriculum adoption on 3rd-5th grade math achievement in North Carolina using a difference-in-differences methodology. We find positive, significant impacts of Eureka Math Squared adoption on math achievement in 3rd and 5th grade (0.11 and 0.15 SDs, respectively) but do not find significant effects in 4th grade or when pooling across all three grade levels. Our pooled point estimate is positive (.07 SDs) but imprecisely estimated. We also find suggestive evidence of positive one-year lagged effects of adoption in 3rd grade, 5th grade, and pooling across all three grade levels, but we urge caution interpreting these results since only two of five districts included in this study adopted Eureka Math Squared early enough to allow estimation of lagged adoption effects.

Research Overview

Research has suggested that high quality instructional materials can positively impact student learning (Chingos & Whitehurst, 2012), and education policy is increasingly incentivizing their use (Council of Chief State School Officers, 2023). Most (but not all; see Blazar et al., 2020) experimental and quasi-experimental studies of mathematics materials suggests different curricula vary in their effectiveness for student learning (Agodini & Harris, 2010; Koedel et al., 2017), underscoring the importance of high quality instructional materials in this subject. In 2012, Great Minds created the well-regarded open curriculum EngageNY, which they updated and revised to Eureka Math in 2013, then to Eureka Math Squared (EM^2) in 2021. EM^2 is designed to build enduring math knowledge in students from pre-kindergarten through Algebra I. This report presents findings from a 2024 efficacy study of the EM^2 curriculum.

In Fall 2024, USC researchers evaluated effects of EM^2 adoption in grades 3-5 using publicly-available school-level end-of-grade math assessment data from North Carolina alongside district-level EM^2 adoption information in that state. The study employs a difference-in-differences design (Callaway & Sant'Anna, 2021) and is designed to meet ESSA Tier II evidence standards.

Data & Methodology

The following variables were included in this analysis.

Outcomes

1. School-level North Carolina end-of-grade scaled score for grades 3-5, 2019-2024

Treatment Variable

1. The year in which each NC district first purchased EM^2, which we infer as that district's school's first year of EM^2 adoption.
   a. Schools in this analysis adopted EM^2 in either 2022-2023 or 2023-2024.
   b. We also infer that schools that did not purchase EM^2 during the study period are not implementing EM^2. EM^2 was first released during the study period, in 2021, so this assumption is a plausible one.

Covariates (all at the school level, collected in each year under study)

1. Enrollment
2. School type (charter vs district)
3. Urbanicity (city vs suburb vs town vs rural)
4. Percent of students eligible for federal free or reduced lunch programs
5. Percent nonwhite students
   a. As of November 2024, the federal database with school-level race/ethnicity and free/reduced lunch data is not yet updated to include the 2023-2024 school year. We produced current estimates by imputing 2022-2023 data for these covariates into the 2023-2024 year, but plan to use 2023-2024 data once they become available. We do not anticipate this decision to affect our estimates of the impact of EM^2 adoption much since it is implausible that the composition of treatment schools changed much relative to the composition of comparison schools between 2023 and 2024.

Additional Variables

1. School-level and district-level identifiers
2. Year (2019-2024, with 2020 omitted due to no outcome data that year)
3. School-level number of students tested, used to weight estimates

Establishing Baseline Equivalence

What Works Clearinghouse evidence guidelines require that studies establish baseline equivalence on key study outcomes between treatment and comparison groups to support causal inference. Using the Hedge's *g* standardized effect size statistic, differences between treatment and comparison groups of less than 0.05 in absolute value are considered to fulfill baseline equivalence, and differences between 0.05 and 0.25 in absolute value are considered to fulfill baseline equivalence if an "acceptable adjustment" is applied to the models.

There are 144 schools in the 5 treated NC districts. We do not have reason to believe that the rest of NC would be a plausible comparison group to these 144 schools, and a naïve comparison of math performance in the final pre-treatment year between treatment and comparison schools bears this out (Table 1, columns 1-3). Therefore, we use pre-treatment outcomes and school-level covariates to restrict the total sample of comparison schools to a more plausible "matched" comparison group for the treated schools.[1] Columns 4-6 of Table 1 show that standardized differences between treatment and comparison groups meet baseline equivalence criteria.

Table 1. Comparison of study outcomes in matched versus unmatched samples

	Unmatched			Matched (demogs + prior scores)		
	C mean	T mean	Hedge's G	C mean	T mean	Hedge's G
G3 Math	546.43	546.18	0.054	546.59	546.56	0.007
G4 Math	546.44	545.14	0.274	546.48	545.61	0.18
G5 Math	545.82	544.74	0.220	545.93	545.35	0.116
Overall Math	546.23	545.36	0.193	546.33	545.85	0.106

In constructing a matched sample, we lose some treated schools alongside a larger number of comparison schools. To the extent that treated schools retained in our matched sample and included in analysis differ from treated schools excluded in matching (i.e., schools with no acceptable match among comparison schools), our estimates of the impact of EM^2 adoption on math performance may apply more closely to the types of treated schools retained in the matched sample. Table 2 compares treated schools retained in the matched sample to treated schools excluded during the matching process.

Table 2. Comparison of treated schools retained and excluded in matching process

	Mean (T schools retained in matched sample, N=107)	Mean (T schools excluded in matching process, N=37)	Significant diff? *p<.05, **p<.01, ***p<.001
G3 Math	545.89	542.99	*
G4 Math	545.04	542.96	

[1] We use 5 nearest neighbor matching with caliper set to 0.25 SDs, following Rosenbaum & Rubin (1985).

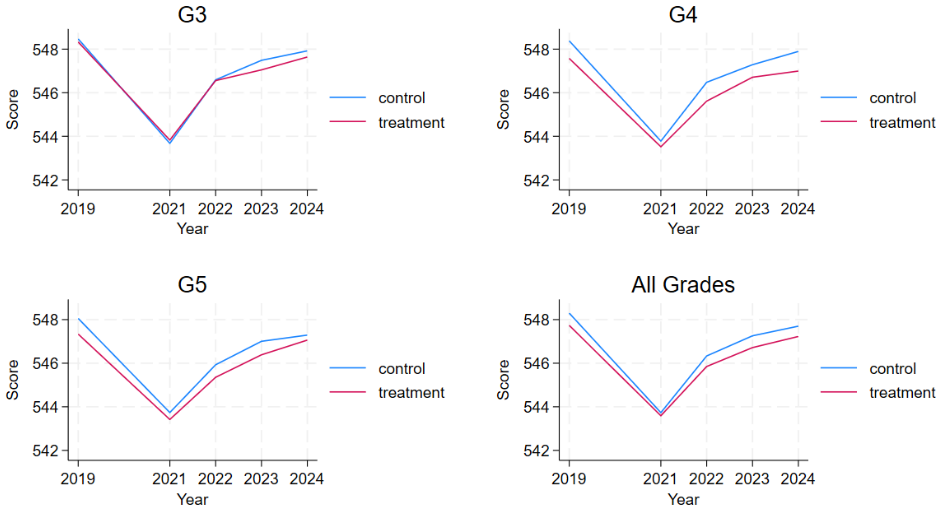
G5 Math	544.47	541.81	
Overall Math	545.11	542.51	*
Enrollment	463.63	453.40	
% Public	99%	100%	
% Town/Rural	17%	70%	***
% FRL	73%	85%	
% Black	34%	36%	
% Hispanic	19%	19%	

Excluded schools are generally comparable to retained schools, with the exception of urbanicity, where excluded schools are more rural. Excluded schools are also very slightly lower achieving on average and in 3rd grade. Therefore, readers should exercise caution when generalizing our impact estimates, especially in rural contexts.

Parallel Trends

To interpret difference-in-difference effects causally, it is important to establish that adopting and non-adopting schools would have experienced similar trajectories in math performance in the absence of EM^2 adoption, conditional on our set of covariates. Where baseline equivalence establishes similar *levels* of math performance prior to treatment, we now establish similar *trends* in math performance prior to treatment.

Figure 1. NC EOG performance over time, by grade (matched sample)



Visual inspection of Figure 1 shows that math performance in adopting versus non-adopting schools follows similar trajectories in pre-treatment years, 2019-2022.

Results

Estimated Effects of Adoption

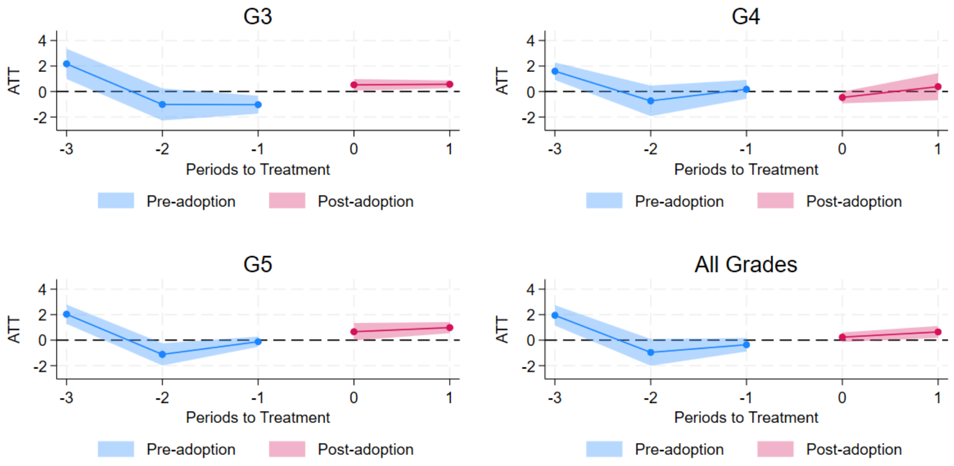
Given that the parallel trends assumption is a reasonable one in this context, we now estimate effects of a school's adoption of the EM^2 curriculum on that school's NC EOG math performance in grades 3-5, conditional on school enrollment, school type, urbanicity, percent of students eligible for free/reduced lunch, and percent of nonwhite students, weighted by the number of test takers in each grade in each school year and with standard errors clustered at the district level. Table 3 displays these estimates numerically and Figure 2 displays them visually, in terms of periods before and after treatment.

Table 3. Estimated impact of EM^2 adoption

Outcome	Estimated Effect (Overall)	Standard Error (Overall)	Estimated Effect (1 year later)	Standard Error (1 year later)
G3 Math	0.536**	0.197	0.567***	0.154
G4 Math	-0.264	0.281	0.383	0.541
G5 Math	0.74*	0.319	0.987***	0.222
All grades (avg)	0.333	0.208	0.646**	0.239

**p*<.05, ***p*<.01, ****p*<.001

Figure 2. Estimated effects of EM2 adoption



Covariates: enrollment, school type, urbanicity, % FRL, % nonwhite
Models use sample matched on pre-treatment outcomes and covariates

We estimate positive, medium-sized effects of EM^2 adoption in 3rd and 5th grade of between 0.5 and 1 scaled score points. We find no significant overall impact of adoption in 4th grade or when averaging across grade levels. However, focusing on effects 1 year after adoption, we find significant impacts of adoption in 3rd, 5th, and when averaging across grades. We urge caution when interpreting this finding because only two districts adopted EM^2 early enough to have data from 1 year after adoption included in this analysis, but it is a promising finding for further investigation in the future. Standardizing the significant overall estimates using standard deviations for each grade level's test, our estimated effects correspond to +0.11 SDs in 3rd grade and +0.15 SDs in 5th grade.

Discussion

We estimated positive effects of the adoption on EM^2 on math performance in 3rd and 5th grade and null effects in 4th grade. However, we are limited in our ability to probe further into potential drivers of these effects and into more complex adoption effects.

For example, because the earliest adoptions in our data occurred in 2022-2023 and we conducted this study in fall 2024, when the most recent outcome data are from spring 2024, we only have 1-2 years of post-adoption data for schools in this dataset. This limits our ability to estimate lagged effects of adoption. Modeling lagged effects would be a useful focus of future analyses, and the little data we do have for districts in their second year of adoption suggest that adoption effects increase from the first to second year. We also have no data on comparison schools' math curriculum. Some may have changed math curricula during the study period and others may not have, potentially diminishing comparability between the treatment and comparison groups.

Our adoption data contained information on the first year that each adopting district purchased EM^2. The possibility that a district could have purchased EM^2 and not implemented the curriculum that year (or at all) injects measurement error into our curriculum adoption measure, potentially inflating standard errors and biasing our estimates toward zero. We also only had access to district-level purchasing data, so the possibility that a district could have purchased EM^2 without every school in the district implementing the curriculum that year (or at all) also injects measurement error into our measure of curriculum adoption, with the same consequences. Finally, we have no data on classroom-level implementation of EM^2. Variation in EM^2 implementation fidelity also injects measurement error into our adoption measure, so could have similar consequences to the above, compared to an error-free measure of adoption. More broadly, this lack of implementation data precludes us from commenting on whether and how fidelity of EM^2 implementation relates to estimated adoption effects.