Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1365

Predicting Persistence and Fadeout Across Multi-Site RCTs of an Early Childhood Mathematics Curriculum Intervention

Tyler W. Watts
Teachers College, Columbia University

Caroline M. Botvin
Teachers College, Columbia University

Drew H. Bailey
University of California, Irvine

Emma R. Hart
Teachers College, Columbia University

Shira Mattera
MDRC

Douglas H. Clements
University of Denver

Julie Sarama
University of Denver

Dale C. Farran
Vanderbilt University

Mark W. Lipsey
Vanderbilt University

This study examined predictors of persistence and fadeout across multiple cluster RCTs that evaluated a preschool mathematics curriculum. We used meta-analytic methods to explore how impacts on student mathematics achievement faded between post-test (i.e., endline) and one-year follow-up. We found that the magnitude of the impact at post-test was a strong predictor of the one-year follow-up impact. Contrary to popular theory, we found that intervention impacts faded faster when students attended sites that showed more learning in the year following the end of the intervention. Factors related to intervention fidelity and dosage did not strongly predict fadeout patterns after considering the magnitude of the post-test effect. Results suggest that educational program evaluators can use immediate impacts to forecast follow-up effects.

VERSION: January 2026

Suggested citation: Watts, Tyler W., Caroline M. Botvin, Drew H. Bailey, Emma R. Hart, Shira Mattera, Douglas H. Clements, Julie Sarama, Dale Farran, and Mark Lipsey. (2026). Predicting Persistence and Fadeout Across Multi-Site RCTs of an Early Childhood Mathematics Curriculum Intervention. (EdWorkingPaper: 25-1365). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/8ajf-4e22

Predicting Persistence and Fadeout Across Multi-Site RCTs of an Early Childhood Mathematics Curriculum Intervention

EEPA-24-OM-5218.R2

Manuscript Received: June 27, 2024
Revised: April 11, 2025; August 2, 2025
Accepted: September 12, 2025

Authors:

Tyler W. Watts (corresponding author)
Teachers College, Columbia University
tww2108@tc.columbia.edu

Caroline M. Botvin
Teachers College, Columbia University

Drew H. Bailey
University of California, Irvine

Emma R. Hart
Teachers College, Columbia University

Shira Mattera
MDRC

Douglas H. Clements
University of Denver

Julie Sarama
University of Denver

Dale C. Farran
Vanderbilt University

Mark W. Lipsey
Vanderbilt University

ORCID iDs:

Caroline M. Botvin: https://orcid.org/0000-0003-2496-8123
Drew H. Bailey: https://orcid.org/0000-0002-7812-1107
Shira Mattera: https://orcid.org/0009-0005-8516-3523
Douglas H. Clements: https://orcid.org/0000-0003-1800-5099

Author Bios:

Tyler W. Watts, PhD, is an Associate Professor of Psychology and Education at Teachers College, Columbia University. His research focuses on the long-term effects of educational interventions.

Caroline M. Botvin, MA, is a doctoral student in the Department of Human Development at Teachers College, Columbia University. She studies early childhood development and the effects of early childhood education.

Drew H. Bailey, PhD, is a Professor in the School of Education at the University of California, Irvine. His research focuses on understanding the developmental processes underlying stability and change in children's academic skills and on the medium- and long-term effects of educational interventions.

Emma R. Hart, PhD, is a postdoctoral research fellow at Boston College. Her research examines when and how educational interventions generate short- and long-run effects.

Shira Mattera, Ph.D., is a senior research associate at MDRC. Her research focuses on evaluation of early care and education programs and practices that support high-quality aligned experiences for children prenatally through age 8.

Douglas H. Clements, PhD, is a Distinguished University Professor, Kennedy Endowed Chair in Early Childhood Learning, and Co-executive Director of the Marsico Institute for Early Learning at the University of Denver. His research and development projects involve learning and teaching early mathematics, curriculum research and development, computer applications, and evaluating research-based curricula and teaching, including taking successful approaches to scale using technologies and learning trajectories.

Julie Sarama, PhD, is a Distinguished University Professor, the Kennedy Endowed Chair in Innovative Learning Technologies, and Co-Executive Director of the Marsico Institute for Early Learning and Literacy at the University of Denver. She conducts research on young children's development of mathematical concepts and competencies, implementation and scale-up of educational reform, professional development models and their influence on student learning, and implementation and effects of software environments (including those she has created) in mathematics classrooms.

Dale C. Farran, PhD, is a Professor Emerita at Vanderbilt University. Her research focuses on the development of young children as it is affected by their educational environments, with a special focus on children from low-income families.

Mark W. Lipsey, PhD, is Professor Emeritus at Vanderbilt University. His research focuses on the effectiveness of interventions for at-risk children and youth and the associated research methods.

Acknowledgments:

We would like to thank Greg Duncan and Robin Jacob for their helpful comments on this work. We would also like to thank members of the Consortium on Early Childhood Intervention Impact for the discussion that contributed to the ideas in this piece. This work relied on several sources of data, each of which was generously supported by grant funding. The research from NYC reported here was made possible by a partnership between Robin Hood, one of the country's leading antipoverty organizations based in New York City, and MDRC. The views and opinions expressed in this work are those of the authors and do not represent the views of the Institute of Education Sciences, the U.S. Department of Education, the National Institutes of Health, nor any of the other funders who supported this work.

Funding:

This analysis was supported by grant funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (1R01HD095930-01A1). Additional funding for the NYC study was provided by the Heising-Simons Foundation, Overdeck Family Foundation, and the Richard W. Goldman Family Foundation. The TRIAD study in Boston and Buffalo was supported by the Institute of Education Sciences (R305K050157). Data for the Tennessee study was provided with support by the Heising-Simons Foundation (2013-26 and 2016-104) and the Institute of Education Sciences (R305K050157 and R305A140126). The EMERGE study in San Diego was supported by the Institute of Education Sciences (R305A080200 and R305A080700).

Declaration of Conflicting Interests:

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

KEYWORDS FROM SCHOLARONE:

Substantive Keywords: Achievement, Early Childhood, Mathematics Education

Methodology Keywords: Meta-Analysis, Multi-site Studies, Longitudinal Studies, Experimental Design

Abstract:

This study examined predictors of persistence and fadeout across multiple cluster RCTs that evaluated a preschool mathematics curriculum. We used meta-analytic methods to explore how impacts on student mathematics achievement faded between post-test (i.e., endline) and one-year follow-up. We found that the magnitude of the impact at post-test was a strong predictor of the one-year follow-up impact. Contrary to popular theory, we found that intervention impacts faded faster when students attended sites that showed more learning in the year following the end of the intervention. Factors related to intervention fidelity and dosage did not strongly predict

fadeout patterns after considering the magnitude of the post-test effect. Results suggest that educational program evaluators can use immediate impacts to forecast follow-up effects.

Researchers and policymakers have taken an interest in early childhood educational (ECE) interventions because of their promised capacity to produce long-lasting changes in children's developmental trajectories (Elango et al., 2016). However, recent literature reviews have found that ECE interventions, like educational interventions at large (Hart et al., 2024), often produce effects that quickly diminish in the years following the intervention (Bailey et al., 2017). Although the field has generated compelling hypotheses that might explain why certain interventions and environments could produce longer-lasting effects (Abenavoli, 2019; Bailey et al., 2017, 2020a), we have produced few empirically supported answers. It remains unclear whether we can identify characteristics of interventions that are likely to produce fading or persisting impacts on student outcomes.

The current study explored multiple theories regarding the fadeout and persistence of intervention impacts to improve our ability to predict when a program will produce more persistent effects. We integrated child-level data collected as part of multiple large-scale cluster-randomized controlled trials (RCTs) of different implementations of the *Building Blocks* (BB) preschool mathematics curriculum. These four scale-up, multi-site RCTs were implemented in five U.S. cities, which included approximately 4,000 children attending publicly-funded preschool and Head Start programs in 175 schools. The studies were designed to assign schools to experimental groups within blocks, with schools assigned to an intervention condition (curriculum implementation) or business-as-usual control (preschool without BB). This design allowed us to generate a dataset akin to a meta-analytic dataset, in which we treated each blocking group (n = 41 blocks) as a unique "experiment."

We began by testing a simple but important question for program evaluation: Do effects measured at the end of the program (i.e., "endline" or "post-test") reliably predict effects

measured at one-year follow-up? If fadeout occurs proportionately to the initial effect size, this suggests that we can generate forecasting models that rely on observed post-test effects to forecast later follow-up effects. Indeed, policy researchers have become concerned with improving forecasting models, as decision-makers often need to understand the longer-term effects of interventions, even when longer-term outcomes will be unavailable (Athey et al., 2019).

We next turned to a host of exploratory analyses that examined whether other observable factors could be used to explain variation in follow-up effects. Here, we examined several influential theories in the evaluation literature (e.g., Bailey et al., 2017; Johnson & Jackson, 2019) about the importance of post-intervention environments for sustaining program effects. We also tested whether intervention fidelity and implementation factors predicted impacts at follow-up, allowing us to evaluate if increasing the contrast between the intervention and the counterfactual group could lead to longer-lasting effects over and above what would be predicted by the post-test effect size alone. Finally, we considered the influence of children's observed characteristics, testing whether the intervention may have produced more persistent effects for children likely to face more disadvantaged environments.

In the next section, we review various theories that have attempted to explain when interventions are likely to produce persistent or fading effects. We then introduce the data taken from the four scale-up studies of BB, and explain the assumptions of our “meta-analytic” approach. In the results section, we report analyses that aggregated effects across the studies, using meta-regression techniques to evaluate how intervention factors forecast follow-up treatment impacts. Finally, we discuss the implications of these results and consider how they might inform intervention development.

Operationalizing "Fadeout" to Understand Connections Between Short- and Long-Term Effects

The terms "fadeout" and "persistence" have taken on a variety of meanings in the educational evaluation literature (see review in Bailey et al., 2020a). In the current study, we use the term "fadeout" to refer to the degree to which treatment impacts following an intervention on a given construct. The term "persistence" describes the degree to which treatment impacts on a given construct remain, or even grow in magnitude, over time. Because these terms describe a longitudinal pattern of mean differences between a treatment and control group on a measure of cognitive or social-emotional development, they can also characterize various learning trajectories over time. For example, in ECE studies, the fadeout of initially positive intervention effects often occurs even though students in both the intervention and control groups continue learning during the post-intervention period. However, if the post-intervention learning rate of children in the intervention group slows down compared with the children in the control group, then the standardized impact estimate between the two groups narrows (Carr et al., 2024). This fadeout pattern, sometimes called "catch-up," unfolds as children in formal educational settings progress rapidly on vertically-scaled measures of cognitive skills during early grades (e.g., Hill et al., 2008). Also relevant for consideration is the possibility of fading or persisting impacts following an intervention with an initially adverse effect. In this case, intervention persistence would be defined as a sustained advantage for the control group (i.e., a continued negative treatment effect), and fadeout would occur if the treatment group outpaced the control group in the post-intervention period. Little attention has been given to this issue in the current literature on fadeout following ECE interventions. However, some work has shown negative impacts following the fadeout of initial positive effects (Durkin et al., 2022).

Operationalizing persistence and fadeout presents several challenges that bear on our ability to predict how impacts for a given intervention will unfold over time. These issues also hold theoretical importance for widely-cited theories of skill building (e.g., Cunha & Heckman, 2007). To illustrate, imagine we implemented a hypothetical reading program in "Setting A." Our evaluation finds a positive impact of 0.50 SDs on reading achievement at post-test (i.e., endline), indicating a strong positive effect of the intervention on the targeted outcome. We follow up with participants one year after the program ended, and we observe an impact of 0.25 SDs on reading, indicating fadeout in the intervening year. Based on these results, what should we forecast for a further follow-up taken two years after the intervention ended? Moreover, if we ran the same intervention in "Setting B" and found an initial effect of 0.20 SDs, what should we predict for follow-up one year later? If fadeout occurs following an absolute pattern of 0.25 SDs per year, as evidence from "Setting A" suggests, then we might predict an impact of 0 at a 2-year follow-up in "Setting A" and an impact of -0.05 SDs at a one-year follow-up in "Setting B."

However, if fadeout occurs *proportionately,* with the pattern of fadeout over time unfolding in relation to the magnitude of the initial effect size, we would generate wholly different predictions. In other words, the results from "Setting A" would suggest a 50% per year decay in impact, implying that the 2-year impact in "Setting A" might be ~0.13 SDs (i.e., 0.50 X 0.25), and the one-year impact in "Setting B" might be 0.10 SDs (i.e., 0.50 X 0.20).

Thus, if we hope to improve our ability to forecast the follow-up effects for a given educational evaluation, we need to better understand whether impacts relate proportionately over time. In the current study, we attempted to address this issue by directly testing the extent to which post-test impacts *predict* the magnitude of follow-up effects. Indeed, most models of skill building theorize that a stronger initial impact on skill attainment will, on average, yield a

stronger later impact on skill development. This assumption is implicit in any regression-based analysis that relates earlier skill measures to later measures (Duncan et al., 2007; Watts et al., 2014), and it is made explicit by prominent theories of skill-building processes (e.g., Cunha & Heckman, 2007). However, current evidence on the fadeout of intervention impacts complicates these predictions (Bailey et al., 2017), calling into question whether exogenously produced impacts on a given skill will relate to skill impacts in later periods.

Prominent Theories Regarding the Factors That May Explain Fadeout

Many theories have been offered to explain why intervention effects fade or persist in the long term (e.g., Abenavoli, 2019; Bailey et al., 2017, 2020a). Such theories can provide program evaluators with useful guides to forecast how program effects may unfold after the end of an intervention. However, empirical research has yet to provide clear evidence evaluating many of the most prominent theories cited in this area. In our exploratory analyses, we examined various additional factors chosen to shed light on these theories and improve our understanding of the conditions under which intervention impacts are likely to persist. Here, we hope to provide a straightforward test of each of these prominent theoretical explanations for fadeout: Do additional observable features of units, treatments, and settings predict follow-up effects over and above the post-test impacts? In other words, do differences between interventions simply produce variation in post-test impact, which will then fade following a monotonic pattern, or do observable characteristics change the pattern of fadeout in important ways?

Post-Intervention Environments

In recent years, significant attention has been devoted to understanding whether variation in the characteristics of the environments children encounter following interventions explains the persistence or fadeout of intervention effects (e.g., Botvin et al., 2024; McCormick et al., 2022;

Pearman et al., 2020). In the program evaluation literature, the concept of "dynamic complementarity" has been particularly influential, as an early intervention might produce larger long-term effects if it is followed by a subsequent high-quality learning experience that capitalizes on previous skill growth (see Johnson & Jackson, 2019). Relatedly, researchers have often hypothesized that fadeout following ECE programs occurs because children in the treatment group do not have sufficient opportunities to build on the knowledge they acquired during a given ECE program. Thus, the learning of the intervention group may slobecause early-grade teachers spend the year teaching these children content that they had already mastered during the intervention (see Engel et al., 2013). This line of thinking, often called the "sustaining environments hypothesis," predicts that a beneficial intervention at time 1 and at time 2 (e.g., pre-k and kindergarten) will positively interact, producing better long-term outcomes for children who received intervention at both time points. Although such predictions make intuitive sense, evidence for a positive interaction between a high-quality ECE intervention and subsequent educational quality has not been consistently observed (Bailey et al., 2020b).

In contrast, other work suggests that early intervention effects may be more likely to persist when children encounter subsequent environments that lead to *lower* levels of learning (e.g., Bierman et al., 2015; Watts et al., 2023). This could happen if a lower-quality environment provides children in the counterfactual condition with fewer opportunities to learn the content provided during the earlier intervention. In other words, if fadeout is driven by control group catch-up, intervention effects may last longer when subsequent instruction provides control group children with fewer opportunities to make learning gains.

In the current study, we examined the influence of post-intervention learning environments by measuring children's learning gains in both the intervention and control groups following the end of the intervention.

Intervention Quality

Although researchers have increasingly turned to examining the subsequent environment when considering the longer-term effects of educational interventions, the programming quality during the intervention period remains an important area of concern. Advocates for ECE have argued that fadeout is likely to result from poor-quality programming, with higher-quality programs more likely to yield larger initial benefits that will last in later periods (e.g., Carolan, 2014).

Because fadeout is inherently concerned with the difference between a treatment and control group on a given outcome, this quality issue can also be understood as reflecting the magnitude of the contrast in experiences between the intervention and control group. Indeed, the best evidence for an early intervention producing large, persistent impacts likely comes from the Abecedarian study (for review, see Elango et al., 2016), which featured an unusually high-quality intervention sustained over several years compared against a counterfactual condition that likely included few supports. In the more recent Head Start Impact Study, several post-hoc analyses have found that post-test intervention impacts were larger for the children who were most likely to stay home during the Head Start year (Kline & Walters, 2016), and another analysis found that impacts were larger and persisted at a higher rate when the treatment group was compared to children in the control group who stayed home (Zhai et al., 2014).

Together, this work suggests that the contrast between the experiences of children in treatment and control group during the intervention period might shape longer-term effects. In

our study, we examined this issue by measuring the treatment and control contrast during the intervention year across several mathematical instructional indicators—all reflecting the magnitude of the change in mathematics instruction introduced by the intervention.

Breadth of Impacts

When attempting to forecast the long-term effects of an intervention, it seems apparent that one should consider the skill domain in question, including what is known about the nature of the typical development of said skill in the absence of intervention. Indeed, Bailey and colleagues (2017) argued that intervention impacts would be more likely to produce long-term effects if they influence malleable, fundamental skills that do not develop quickly without intervention.

In the case of mathematics achievement, fadeout might be expected in the short term. Although preschool mathematical skills are fundamental and malleable, children who did not receive the intervention are poised to quickly learn early mathematics content when they first encounter math instruction upon entering kindergarten. However, theory also suggests that mathematical skill development is foundational for learning in other skill domains (e.g., Clements & Sarama, 2011), opening the possibility that math learning could transfer to broader capacities not directly taught to counterfactual children after the intervention ends. In the current setting, if the BB program also encouraged the development of, say, executive functioning, then transfer effects from executive functioning to mathematics achievement may lead to more observed treatment impact persistence on mathematics in the long-term (e.g., Clark et al., 2010; Clements et al., 2020). In the current study, we evaluate this possibility by testing whether the impacts of the ECE interventions on *non-mathematical* cognitive skills (i.e., non-targeted skills) lead to more persistent effects on mathematical skills at follow-up.

Demographic Characteristics of the Children Served

Finally, when considering how the characteristics of other environments that children encounter might affect intervention fadeout and persistence, broader socioeconomic influences have also received substantial attention in the literature. Indeed, studies evaluating ECE programs often find that program impacts are larger for children who grow up in more disadvantaged environments (see review in Watts et al., 2023). Previous evidence leads us to make clear predictions regarding the role of demographic characteristics in shaping persistence and fadeout trajectories. Because English language learners and socioeconomically disadvantaged children may receive fewer high-quality post-intervention learning opportunities on average, more persistent impacts may be observed for these groups, consistent with findings from other ECE studies (Bitler et al., 2014; Duncan & Sojourner, 2013; Watts et al., 2023).

Current Study

The current study relies on data from four scale-up interventions of an early childhood mathematics curriculum. Although these evaluations differed in implementation factors and design elements, all four studies featured random assignment to an intervention that included the implementation of the BB curriculum. The curriculum was developed by Clements and Sarama (2008) to overhaul how mathematics is taught in public preschool programs. Previous work has described the curriculum extensively (Clements, Sarama, Spitler, et al., 2011). Here, we briefly describe some of the most salient features of the program.

The BB curriculum used developmental research and theory to help children learn mathematical concepts and procedures through specific learning trajectories. These learning trajectories are centered on mathematical goals (e.g., counting), and students learn how to progress from the basic level of that goal (e.g., verbal counting) to more advanced understanding

(e.g., cardinality, application of counting strategies to solve problems). The BB approach involves significant teacher professional development, during which teachers are encouraged to "mathematize" children's everyday activities and play. The curriculum depends on small group interactions during intentional activities, encouraging child language development and play-based learning. It also integrates technology (i.e., the curriculum includes companion software) and formative assessment into classroom procedures.

The BB curriculum has been researched extensively through multiple scale-up studies, and it is currently being used in several large-scale public preschool programs (e.g., Boston; New York City). For the current study, we synthesized data across four of these large scale-up evaluations: 1) the TRIAD evaluation implemented in Buffalo, NY and Boston, MA (Clements, Sarama, Spitler, et al., 2011); 2) the Vanderbilt Peabody College evaluation of BB from Tennessee (Hofer et al., 2013a; 2013b); 3) the EMERGE scale-up from San Diego, CA (Clements et al., 2020); and, 4) the Making Pre-K Count (MPC) scale-up implemented by MDRC in New York City, NY (Morris et al., 2016). Importantly, treatment effects varied considerably across these evaluations. The TRIAD study in Buffalo and Boston featured an extensive coaching and professional development model, along with the curriculum, and researchers reported a large experimental impact on end-of-preschool mathematics achievement. Clements, Sarama, Spitler, et al. (2011) observed a standardized effect of 0.71 that faded by the end of first grade (g = 0.28; Clements et al., 2013). A parallel evaluation conducted in Tennessee also found a large intervention effect at the end of preschool (d = 0.58), though this effect faded to non-significance in kindergarten and first grade (Hofer et al., 2013a). Finally, the evaluations in San Diego (Clements et al., 2020) and New York City (Morris et al., 2016) both reported

smaller, non-statistically significant effects of the program on mathematics test scores at the end of preschool.

All four evaluations of the BB curriculum relied on cluster-RCT designs, in which schools (or preschool/Head Start centers) were clustered into blocking groups based on common site characteristics and then randomized to experimental groups (one study used some classroom-level randomization—an issue we revisit below). In our analysis, we treat each block as a mini-experiment that can be analyzed similarly to meta-analytic data. Therefore, our key dependent variable is the block-level standardized impact of the program on children's mathematics scores at one-year follow-up (i.e., end of kindergarten), and our key independent variable is the block-level impact of the program on post-test scores (i.e., end of preschool). Importantly, we consider the estimates for each block to be akin to quasi-experimental impacts of the curriculum program. Though the original evaluations used random assignment (with some departures described below), each respective block contains few schools, making the assumptions of random assignment tenuous for internal validity. However, the average of these block-level impacts should equal the unbiased average treatment effect across the multi-site trials, and our inclusion of child pre-test data heavily strengthens the internal validity and reliability of our estimates.

With our approach, we attempt to address the following questions:

1. Do initial impacts observed at post-test predict impacts at one-year follow-up?
2. Do other observable factors, like the level of post-intervention gains for the treatment and control group, the impacts on other skills, treatment quality and fidelity indicators, and demographic characteristics of the children served, provide additional predictive validity for forecasting follow-up impacts over and above the post-test impact?

Method

The supplementary file provides further details. Here, we attend to key study features.

Data

We examined patterns of fadeout using data obtained from four RCTs of the BB preschool mathematics curriculum intervention using integrative data analysis techniques. These four longitudinal evaluations were implemented across five study sites: Boston and Buffalo (Clements, Sarama, Spitler, et al., 2011), Tennessee (Hofer et el., 2013a), San Diego (Clements et al., 2020), and New York City (Morris et al., 2016). All trials were designed to examine the effects of BB on children's achievement at the intervention's end (i.e., spring of preschool) and at one-year follow-up (i.e., spring of kindergarten). Data were cleaned at the study level before merging across sites.

Within sites, school-level randomization occurred within blocking groups in which schools were matched on common characteristics. This matching process differed across the various trials. For Buffalo and Boston, participating public schools offering pre-k services were ranked on elementary school achievement scores within district to create the blocks. In Tennessee, schools were blocked based on school type (Head Start vs. public school) and the number of participating classrooms within a school. In San Diego, schools were blocked by district and classroom care type: full- or part-day[1]. Finally, in New York City, schools were blocked by borough, school type (community-based versus school-based), and whether the school served predominately Hispanic children.

We began by merging student-level data across the five sites[2]. We then reduced the sample for the current analysis in several noteworthy ways. First, study sites in Boston, Buffalo, San Diego, and New York City evaluated the inclusion of additional treatment components, and units randomized to these groups were removed. Consequently, our measured treatment impacts

represent the comparison of the BB program during preschool only (i.e., no intervention extensions into kindergarten) versus "business-as-usual" control. Note that, unlike many other studies often cited in the ECE fadeout literature, the control group in each of these studies still attended preschool.

To make the treatment and control comparison as consistent as possible across blocks, we removed schools/classes assigned to a self-regulation intervention component in San Diego (child n = 392), schools assigned to the "follow-through" BB intervention condition in Buffalo and Boston (child n = 471), and children assigned to a kindergarten afterschool tutoring program in New York City (child n = 173). Next, we limited the analytic sample to blocks containing valid observations for treatment and control group schools[3]. This resulted in the inclusion of 41 blocks and 3,381 participants across 166 schools (mean of ~4 schools per block) with at least one valid math assessment. Among this final sample, 1,680 children were in the treatment group, and 1,701 were in the control group.

Measures

Short- and Medium-Run Treatment Impacts on Math

Our key analytic approach relies on generating block-level treatment impacts on children's mathematics achievement. We begin by describing the mathematics measures available in the data and the transformations we used to make them comparable across the sites before describing how the respective impact estimates were generated.

Children's math performance was assessed three times: (1) at the beginning of preschool (pre-test); (2) at the end of preschool (post-test/endline); and (3) at the end of kindergarten (one-year follow-up). Math achievement was measured using various assessments across the five sites. Although some sites collected a single measure of math performance across assessment

waves, others collected multiple measures. For our analyses, we prioritized measures that had overlap across the sites. Though the assessments were generally consistent within sites for the entire study period, there were several cases where the measures of math performance changed from one assessment to the next. Overall, the math assessments collected across sites and included in our analyses included: the *Woodcock Johnson (III) - Applied Problems* (WJ-AP; Woodcock et al., 2001), the *Research-based Elementary Math Assessment* (REMA; Clements et al., 2008), the *Tools for Early Assessment of Mathematics* (TEAM; Clements, Sarama, & Liu, 2008; Clements, Sarama, & Wolfe, 2011b), and a measure of math performance from the *Early Childhood Longitudinal Study-Birth Cohort* (ECLS-B; Najarian et al., 2010).

Due to the measurement differences across sites, we generated standardized mathematics composite scores for each site and wave using the available math tests. We followed the same process at pre-test, post-test, and follow-up for the math measures available within a given site (see Table A1 in the supplementary material for the list of measures collected across each site at each wave). First, we standardized each respective measure at a given timepoint by subtracting each child's score from the mean within the site and dividing by the site standard deviation for the control group. If a given site had more than one math assessment administered at a given timepoint (e.g., Tennessee had both the REMA and WJ *Applied Problems* administered at post-test), the two standardized scores were averaged together, and the resulting composite was re-standardized using the site-level control group standard deviation. Thus, the final standardized composites of mathematics achievement removed between-site variation due to measurement differences across the 4 scale-up studies, but they allowed for between-block variation within a given site.

We relied on these standardized composite variables because our data did not contain consistent mathematics measures across all five sites. Appendix Table A2 provides correlations among each respective mathematics measure administered at each site. As this table reflects, correlations among the various pre-k post scores and kindergarten follow-up scores were quite high and fairly consistent across the various measures, ranging from 0.61 to 0.75. Although we would have preferred to have a single measure of math that could be compared across sites, our composite approach can be likened to most meta-analyses, as the underlying studies in a given meta-analysis typically do not contain a single common outcome measure. To examine whether our results were sensitive to the measurement particularities of any single site, we examined sensitivity tests that sequentially dropped each site from the analyses (results described in more detail below).

Next, we used these composite mathematics measures to generate child-level treatment impact estimates for each respective block at post-test and one-year follow-up. To generate these estimates, we ran OLS regression models for each block, in which the post-test composite score was regressed on treatment status and controls for gender, race, ELL status, and student age at pre-test. Importantly, we also controlled for pre-test math composite scores. The resulting treatment impact estimates and standard errors were then saved for each respective block, and the process was repeated using the follow-up composite scores of kindergarten mathematics achievement. In these regressions, any missing scores on student-level controls were imputed using block-level means, and dummy variables were included to account for this missing data procedure (see Appendix Table A3 for descriptions of missing data on covariates).

As we describe below, we used meta-analytic techniques to analyze relations between these block-level impact estimates, with blocks being akin to “studies” in a meta-analysis.

However, because the math measures were standardized at the site-level, our effect sizes were different than what would be included in a typical meta-analysis because they included between-block variation within each site. Further, as mentioned above, we consider these block-level estimates to be akin to quasi-experimental impacts of the intervention. Indeed, we used child-level analyses to estimate these impacts despite the fact that schools were randomly assigned, because so few schools were included in each block. On average, the blocks had 77 children observed at post-test (range of 23 to 291) and 70 at follow-up (range of 22 to 251), with roughly 4 schools (range 2 to 29) per block (including classrooms that were randomly assigned in San Diego[1]). The average treatment impact standard error produced by these child-level regressions at post-test was 0.21, and it was 0.24 at follow-up.

Because of the school-level clustering, we scaled the standard errors for each block using a variance inflation factor that included the observed ICC from each block. Yet, the assumption that school-level random assignment will produce groups equal on expectation with few schools randomly assigned within each block is tenuous at best, making the inclusion of the child pretest composite scores crucial for obtaining unbiased estimates at the block level.

Predictors of Follow-Up Impacts

Post-Intervention Learning Gains. To measure learning gains following the end of the intervention, we used a different measurement process than the aforementioned process for generating treatment effect sizes. For these variables, we relied on a consistent math measure available at *both* the post-test and follow-up time points. REMA performance at post-test and follow-up was used in the Buffalo, Boston, and Tennessee sites. For the San Diego site, we included post-test and follow-up performance on the TEAM, while performance on the WJ-AP was used for the NYC site.

We began with a student-level measure of gains in the control group. For this measure, non-standardized post-test math performance was subtracted from non-standardized follow-up math performance for each control-group participant, and this difference was then divided by the average site-level control-group post-test standard deviation. This was similar to the standardization process for the mathematics outcome measures described above but more flexible as we did not center observed gains within sites. The measure of gains for the treatment group was similarly generated within each site by calculating the difference in non-standardized post-test and follow-up math performance for each treatment-group participant and dividing the resulting value by the site-level *control-group* post-test standard deviation (i.e., this ensures that treatment and control group gains are scaled using the same factor in each site). In our analytic models, this measure of gains was accompanied by a measure of average post-test math performance in each group to account for the likelihood that gains from post-test to follow-up varied by post-test performance levels. Here, we simply took the block-level mean for both the treatment group and control groups, respectively, on the standardized composite math scores described above.

Finally, these respective treatment- and control-group measures were aggregated to the block level, such that each block had a unique measure of treatment gains, control gains, treatment post-test level, and control post-test level. To generate the measure of *total* post-test gains, we averaged the two gains measures together at the block level, ensuring that treatment and control group gains were equally weighted within blocks. Similarly, the total post-test level was created by averaging the block-level treatment- and control-group post-test level values.

Demographic Characteristics. We examined whether two participant demographic characteristics, aggregated at the block level, predicted medium-run treatment impacts. This

included the proportion of Black, non-Hispanic participants and the proportion of English Language Learner participants within each block. These measures were generated by determining the average proportion of children that fell into the demographic characteristics within each block. For our analyses, these measures were rescaled to 10-percentage-point units to improve the interpretability of our results.

Short-Run Impacts on Non-Math Skills. Short-run treatment impacts on measures of non-math skills were calculated with the same approach as was employed to calculate math treatment impacts. A post-test composite of non-math skills was created within site by standardizing, averaging, and re-standardizing performance on a variety of non-math measures including measures of language and literacy (e.g., Woodcock-Johnson Letter Word) and measures of executive function (e.g., Head-Toes-Knees-Shoulders). Supplementary Table A1 presents the various measures of non-math cognitive skills included in each site. Additional information about these measures can be found in the online supplemental material. Average treatment impacts were estimated using the same process described above for math achievement.

Preschool Classroom Instructional Characteristics. Classroom instructional quality was measured using the *Classroom Observation of Early Mathematics Environment and Teaching* (COEMET) assessment, a three-hour classroom observational measure. During the spring of preschool, trained observers used a series of 28 Likert-scaled items to rate the quality of the children's preschool math environment[4]. We relied on three subscales of the COEMET, including the quality rating of the specific mathematics activities (SMA), the total number of SMAs observed, and the total time spent on math instruction during the observational period. A specific math activity (SMA) was defined as an activity that was conducted intentionally by the teacher and (a) involved several interactions with one or more students, or (b) was designed to

develop knowledge of math. Assessors completed a rating form for each SMA to determine the quality of math instruction related to that activity (e.g., "the teacher began by engaging and focusing children's mathematical thinking"). Although the SMA form was completed each time an SMA was observed, both the total number of math activities and the total time spent on math were recorded once to reflect the assessors' overall impression of the learning environment. We define the total number of math activities as the total number of full and mini SMAs. Mini SMAs were meant to capture activities that incorporated mathematical concepts but were either short in duration or did not include teacher involvement. Additionally, our measure of time is best understood as the sum of the average proportion of time spent on SMAs relative to the duration of classroom observation. It should be noted that while all study sites implemented some version of the COEMET, there were several differences in the measure across sites, which are described in more detail in the online supplement.

Because the COEMET was designed to capture the teaching practices encouraged by the BB program, we ultimately used our three COEMET indicators (overall COEMET quality score, total number of math activities, and total time spent on math activities) to measure block-level treatment impacts on mathematics instruction. Thus, as with the measures of mathematics achievement, the COEMET instructional indicators were standardized to the control group SD within sites to account for differences in measurement across sites. We then regressed these standardized composites on the treatment indicator at the child level within each block to derive block-level treatment impacts on measures of mathematics instruction. Because this approach again used child-level impact estimates to examine treatment impacts on classroom-level factors, we revisit this choice with sensitivity analyses described below.

Design Inconsistencies and Limitations

In this integrative analysis, we attempted to align the data across the five sites in a consistent fashion to allow for meta-regressions that would examine patterns of fadeout and persistence across the included studies. However, each site contained idiosyncrasies that had to be handled individually, including measurement and design decisions that are described in more detail in the supplemental file. Given these inconsistencies and design limitations, we employ a bevy of sensitivity tests, with the most notable test including a specification in which we exclude each of the five study sites, in turn, to examine whether our estimates were strongly affected by the design complications of a single site.

A few important design limitations are worth noting here. First, the TRIAD study (implemented in Boston and Buffalo) suffered from some compromises to random assignment, as six schools switched treatment groups after random assignment, three schools dropped out, and four schools were added. In the preferred models, we use the "final" treatment designation for each included school (consistent with previous reports; Clements, Sarama, Spitler, et al., 2011), though we also examined models that dropped all compromised schools, and treatment impacts were consistent with our preferred estimates. In Boston, study administrators were also later concerned that the blocks were not properly used during random assignment, as several blocks had uneven assignment (i.e., multiple or all schools assigned to the same condition within a block). The data provided by the original study administrators to the current analytic team includes the blocking group matches that were designated based on school-level achievement, which represents the intended method for pre-random assignment school matching. As we noted earlier, our preferred models only include blocks that contained valid observations from both treatment and control schools (i.e., within block variation in treatment). However, given the concerns over whether the blocking procedure was actually followed in Boston, we also

examined models that completely dropped Boston blocks from our key estimates (described in more detail below).

In San Diego, different randomization procedures were followed depending on the number of participating classrooms within each school. Study schools with one participating classroom were stratified based on full-day or half-day status within each district, creating five randomization blocks (see Clements et al., 2020). However, randomization occurred at the classroom level for schools that had two participating classrooms. For the purpose of our analyses, we essentially ignore this discrepancy as we generate impacts at the child level with no additional stratification based on school. Study researchers in San Diego were also unable to collect individual-level demographic information for each child from the district, leaving the majority (76%; see Table A3) of children missing data on demographic characteristics (i.e., race, ethnicity, and English proficiency). For the key demographic features used in our analysis, we simply took the average of the non-missing observations in each block, and assessed the sensitivity of our findings to this issue in San Diego by dropping the San Diego site completely from our analyses (shown in the supplemental file).

Finally, both San Diego and New York City administered potentially compromised pre-tests, as study administrators were unable to finish pre-test data collection until partway through the fall semester of the school year. This could mean that our estimates of post-test and follow-up treatment impacts are biased downward for these sites due to our inclusion of the pre-test as a covariate at the child level. In analyses described below, we also examined models that did not use the pre-test as a control when calculating block-level post-test and follow-up impacts.

Analytic Plan

Main Analyses

To test the extent to which short-run, end-of-treatment impacts persisted a year after the intervention ended, we executed a series of regression analyses to estimate persistence rates. Random-effects meta-regressions were conducted in R using the "metafor" package (Viechtbauer, 2010). Effects were weighted using the standard weighting procedure for meta-regression with metafor, which relies on the inverse variance-covariance matrix of the meta-regression model (i.e., the SE of each effect is taken into account with between-study variability). The following model was used to estimate patterns of persistence across the four BB studies:

$$Impact_{fbs} = \beta_{0s} + \beta_1 Impact_{pbs} + \ \varepsilon_{bs}$$

$$\beta_{0s} = \gamma_{00} + \tau_{0s}$$

where $Impact_{fbs}$ represents the follow-up (i.e., spring of kindergarten) impact for a given block, *b* in site *s,* $Impact_{pbs}$ represents the post-test (i.e., endline) impact for the same block. Here, the key coefficient is represented by β_1, which captures the predictive relation between post-test impact and follow-up impact in effect size units. This can be understood as the *conditional persistence rate* between post-test and follow-up (e.g., a coefficient of "1" would suggest that a follow-up impact of .25 SD would be expected given a post-test impact of .25 SD).

In this model, β_{0s} is allowed to vary across sites, with the cross-site average intercept, γ_{00}, indicating the predicted follow-up effect when a given post-test impact is zero. Thus, γ_{00} is thought to capture treatment-driven "unmeasured mediators" that lead to impacts at follow-up that are not captured by post-test impacts (see Hart et al., 2024 for additional discussion of these model parameters). We estimate the variance in τ_{0s}, which we interpret as the amount of variance in follow-up effects that remains once the post-test impact is taken into account.

We then further examined the extent to which various factors predicted follow-up treatment impacts on math skills beyond post-test treatment impacts on math skills. For these exploratory analyses, we added each block-level characteristic of interest (i.e., measures of learning gains for the treatment and control group, block-level demographics, impact on non-math skills, and various COEMET factors of implementation) to the model described above.

Results

Descriptive Results

Table 1 presents participant demographic characteristics across the five study sites (see Appendix Table A4 for differences by treatment condition). As Table 1 reflects, participant race varied considerably across the sites, with non-Hispanic Black students comprising 56% of the sample in Buffalo, 40% of the sample in Boston, 77% of the sample in Tennessee, 5% of the sample in San Diego, and 35% of the sample in New York City. Boston, San Diego, and New York City each had larger percentages of Hispanic children (40%, 48%, and 58%, respectively), also reflected in the percentage of students qualifying as limited English proficient.

Table 2 presents descriptive information at the block level for the key variables included in our analyses. As Table 2 shows, the pre-test adjusted post-test impacts were largest for Buffalo ($M = 0.67$) and smallest for NYC ($M = 0.00$). Averaging across all blocks, the pre-test adjusted impact at post-test was 0.26 ($p < 0.05$) and the follow-up impact was 0.10 ($p = 0.15$)[5]. These estimates were generated by using the random effects meta-regression model and fitting the post-test impact and follow-up impact, respectively, with no predictors. Not surprisingly, Table 2 also shows that large impacts were observed across the sites on instructional characteristics, including COEMET quality, time on mathematics instruction, and the number of math activities taught.

In the supplemental file, we also provide additional descriptive information at the block level, and include figures showing treatment and control group trajectories on the math tests administered for each blocking group within each site. These figures graphically represent the learning gains measures that we used to examine post-intervention trajectories across the various blocks. Several patterns are worth noting in these figures. First, the trajectories make apparent the need for pretest adjustments when deriving impact estimates, as the treatment and control groups were often at different levels on the pretest, despite the randomization of schools. Second, the figures generally show a steady growth rate for the control group through both pre-k and kindergarten, with the treatment group showing an accelerated growth rate during pre-k, followed by a slower growth rate during kindergarten. These general patterns are consistent with the idea that the intervention provided a short-term acceleration of learning for the treatment group, but this learning rate fell during kindergarten.

Supplementary Table A5 provides correlations among our key variables, all aggregated to the block level. Of note, the unadjusted correlation between the post-test impact and follow-up impact was 0.56 (p <0.001), suggesting covariation between impacts at both timepoints. Also of note, gains in the control group had virtually no correlation with post-test impact (0.05), indicating that factors facilitating learning in the post-intervention period did not exert a large effect on the treatment group during the treatment period. In contrast, the post-intervention gains in the treatment group were strongly and negatively related to post-test impact magnitude (-0.58, $p < 0.001$), suggesting that treatment children who were in highly effective *BB* sites were more likely to show slower math learning rates after the intervention ended.

Finally, in the supplemental file, we also provide a forest plot that shows post-test and follow-up impacts on our composite math measure for each blocking group.

Regression Results

Figure 1 plots the observed block-level post-test impact estimates against the follow-up impact estimates, providing our key results. We observed a slope of 0.40 (SE = 0.13, $p < 0.01$; see Table 3), suggesting that, while follow-up impacts overall fade somewhat, the extent to which they persist is proportionate to the size of the post-test impact across the distribution of post-test effects. Interestingly, this basic linear pattern of impact decay between post-test and one-year-follow-up was evident even for the few blocking groups that had *negative* post-test impacts, suggesting that impact fadeout may also be expected when interventions produce negative short-run effects. Importantly, we observed that the regression line nearly crosses through the origin, indicating a y-intercept term near zero. This suggests that any block that produced a zero SD impact at post-test was also predicted to have a zero effect at follow-up.

In Table 3, we examined the extent to which gains in the follow-up period within- and across- intervention groups predicted follow-up impacts when controlling for the post-test effect size. In other words, these models test whether post-intervention learning provides predictive validity for follow-up impacts over and above post-test impacts alone. As Column 2 reflects, total learning gain during the post-intervention period was negatively predictive of the follow-up effect (β = -0.45, SE = 0.24) when controlling for the post-test impact, though this coefficient was imprecise and not statistically significant ($p < 0.10$). When the post-intervention gain measure was disaggregated by intervention group, we found that control group post-intervention gains were strongly negatively predictive of follow-up effects (β = -0.53, SE = 0.20, $p < 0.01$). The coefficient for post-intervention treatment gains was near zero and not statistically significant. This pattern, consistent with the figures in the supplement, indicates that given the strong correspondence between the post-test impact and post-intervention gains rate in the

treatment group, it is the amount of post-intervention learning in the control group, unaffected by post-test impact, that has the most influence on the follow-up effect.

When both treatment- and control-group gains were included in the model together (Column 5), control-group gains (β = -0.91, SE = 0.27, $p < 0.001$) again produced a larger coefficient than treatment gains (β = 0.67, SE = 0.34, $p < 0.05$). Here, the predictive validity of the post-test impact was attenuated by the post-intervention measures. However, including all of these predictors together did cause SEs to increase substantially, suggesting that the model is likely to be over-specified due to a set of highly correlated predictors.

Table 4 presents estimates for other exploratory predictors measured at the block level[6]. Across these columns, we largely found that these characteristics were not strongly predictive of further follow-up effects after accounting for the post-test impact. However, several findings are noteworthy. First, contrary to our expectation, the impact on non-math cognitive skills, conditional on the post-test effect, was *negatively* predictive of follow-up impact (Column 1), though this estimate was not statistically significant (β = -0.20, SE = 0.16). In general, we found little indication that effects on instruction were strongly predictive of follow-up effects after accounting for the post-test impact, indicating that the effect of implementation factors on later follow-up impact was likely fully accounted for by the intervention impact at the post-test (see Table A5 for the full list of correlations between each predictor and post-test impact). However, the impact for time spent on mathematics was negatively predictive of later follow-up impact (β = -0.06, SE = 0.03, $p < 0.05$). Finally, we did not observe statistically significant predictions for the two demographic characteristics considered: percent limited English proficiency and percent non-Hispanic Black.

Sensitivity Analyses

Several sensitivity analyses were conducted to examine the robustness of our main results. First, we assessed whether differences in site-level characteristics may have biased the observed patterns of persistence/fadeout by replacing the random effect in our main model with a site-level econometric fixed effect. Introducing this site fixed effect produced estimates consistent with our primary results, indicating that unobserved characteristics between sites did not appear to bias our estimates (see Table A6 and A7). Next, we removed the pre-test adjustments from our key variables and regression weights to account for the inconsistency in pre-test timing (which could deflate post-test and follow-up impact estimates). Again, results were largely consistent with those reported in the main text (see Table A8 and A9).

We also assessed the extent to which design complications may have impacted our results by examining our main analyses with each site sequentially excluded. Overall, the estimates remained relatively consistent in magnitude and significance across each series of regressions (the post-test slope coefficient ranged from 0.28 to 0.54; it was marginally significant ($p < 0.10$) when NYC was dropped), indicating that the noted issues with randomization, blocking, and pre-test data collection at various sites did not substantially affect our findings (Tables A10 and A11). Notably, dropping the Boston site, which had the most substantial issues with the block variables, had a negligible effect on our key result. We also examined models that dropped schools from Boston and Buffalo that either switched treatment groups or were added after initial random assignment (Tables A12 and A13). Again, effects were similar to our main results.

Tables A14 and A15 examine the extent to which our estimate of the conditional persistence rate was influenced by blocks that produced negative impacts at post-test. Dropping these negative post-test impacts limited our model to 31 blocks. However, the conditional persistence rate between post-test and follow-up impact remained similar: 0.34 *(*$p < 0.10$).

In Tables A16 and A17, we present results that used the robust variance estimator to further adjust estimates for the site-based clustering of the data (i.e., the 41 block estimates were derived from 5 sites). Although this estimation technique is common for clustered data in meta-analyses, with only 5 sites in our data for clustering, it is likely the case that we do not have enough higher-order units to draw unbiased standard errors. With this adjustment applied, the SE for post-test impact increased from 0.13 to 0.17, and the estimate was no longer statistically significant with no other controls in the model.

Because we generated child-level impact estimates for classroom-level measures of COEMET-based variables, we also included a sensitivity test in which we took the average scores on each variable for both the treatment and control group and then calculated the difference in these scores at the block level. These simple "difference" variables were used in Table A18. Here, we again observed largely null effects for these measures. The coefficient for time spent on math was negative in direction, but not statistically significant.

Finally, we included several additional tests to examine the influence of pre-test differences on our forecasting results, and we found little evidence that our key models were heavily influenced by endogenous differences between children that were present at pre-test. Indeed, pre-test "impacts" (i.e., treatment and control differences on the pre-test at the block-level) were not predictive of follow-up effects, and we also found that our key results were consistent when controlling for pre-test impacts. We also examined the effect of differential attrition on our results by excluding any block that had a differential attrition rate higher than 10 percentage points (i.e., the difference in the probability of leaving the sample between treatment and control was higher than 10 p.p.), and again found no indication that differential attrition heavily influenced findings here.

Discussion

The current paper used an integrative analysis of four scale-up studies of the *Building Blocks* early mathematics curricular intervention for preschool-aged children to examine specific hypotheses regarding our ability to forecast follow-up effects following the end of an intervention. Although the field of ECE research has become intensely focused on understanding the conditions under which intervention effects are likely to persist or fade, few studies have generated empirically-supported findings that could improve the field's ability to predict and interpret follow-up impacts following the end of an intervention. Across five sites, we observed an average treatment impact of 0.26 SDs at the end of preschool, and this impact fell to 0.10 by the end of kindergarten. As predicted, this average pattern of persistence reflected a proportional relation between post-test impacts and follow-up impacts at the block level. The post-test impact was a strong predictor of follow-up impact one year later, with a slope term of 0.40, suggesting a conditional persistence rate of approximately 40%. In this model, we observed a y-intercept near zero, indicating that follow-up impacts were not observed for blocks that did not also produce short-term impacts.

We then attempted to predict the remaining heterogeneity in follow-up impacts with a variety of block-aggregated measures designed to assess prominent theories regarding fadeout. These factors included treatment impacts on non-math skills, math learning in the post-intervention period, and the fidelity/quality of treatment experience. Despite the large number of schools and children included in the current study, estimates were generally imprecise, and only a few were statistically significant. Among these additional predictors, we found that the amount of math learning by children in the control group after the intervention ended was the single best predictor of follow-up impacts after controlling for the post-test effect (more control group

learning predicted faster fadeout relative to the treatment group fadeout, which was largely proportionate to the post-test impact). Below, we interpret estimates in terms of their effect sizes, but it should be noted that a large degree of uncertainty was associated with these estimates.

Forecasting Follow-Up Effects Using Short-Term Effects

Our results suggest that if program evaluators hope to forecast the follow-up effects of their interventions, their post-test effect sizes on the same outcomes likely provide the best place to start. This may seem intuitive, and to some degree, this ~40% persistence rate by one-year follow-up is implied in other meta-analyses (e.g., Li et al., 2020). However, most previous work did not align measures and impacts over time in a way that would allow for the examination of the rank-order stability of impacts at follow-up. As we described above, understanding these relations is a critical step to improving our forecasting capabilities when longer-term follow-up is not possible.

However, generalizations to other evaluation settings should be made with caution. When drawing comparisons to other evaluations, it should be noted that our study included a relatively narrow comparison between the treatment and control groups when compared with broader evaluations that test multi-faceted educational programs versus a control condition (e.g., the Head Start Impact Study, a state pre-k evaluation). Here, we evaluated the impact of an intensive curricular intervention on the same construct (mathematics achievement) over time, with pre-k business-as-usual control groups and a follow-up period extended to only one year. Future studies should seek to replicate these findings in other intervention settings featuring different contrasts between treatment and control groups. Indeed, one could imagine these results differing for interventions targeting other developmental periods or skills. Yet, it should be noticed that a recent meta-analysis of highly diverse educational interventions targeting cognitive and social-

emotional skills found a strikingly similar relation between post-test impacts and one-year follow-up effects for both cognitive and social-emotional skills when compared to the one-year persistence rate reported here (Hart et al., 2024).

Perhaps surprisingly, we found that the general model of impact decay was also valid for the few blocks that produced *negative* post-test impacts on mathematics achievement, suggesting that unintended negative effects also faded substantially in the year following the intervention. Importantly, we found that our overall pattern of results was robust to the inclusion or exclusion of blocks that had negative post-test effects, but the consistency of this pattern is notable given that theoretical considerations of fadeout have largely centered on the pattern of diminishing impacts following positive short-term boosts to achievement (e.g., Bailey et al., 2017). However, the pattern of findings makes sense in light of explanations of skill development that posit the importance of many dynamic factors that are not immediately affected by the intervention, but are likely to help push (or pull) children back toward their previous trajectories in the years following the treatment (e.g., Bailey et al., 2016).

For prominent theories of skill building in the evaluation literature (Cunha & Heckman, 2007), our results should be seen as providing partially confirmatory evidence. On the one hand, larger skill impacts at time 1 did lead to larger skill impacts at time 2, suggesting some transfer of skill attainment in mathematics achievement after the intervention ended. Moreover, the null effect observed for the y-intercept in our model suggested that, in the case of the *Building Blocks* intervention, positive follow-up impacts were *only* observed for blocks that also produced positive effects at the end of the intervention. This again suggests that any longer-term intervention effects were likely due to math gains made during the initial intervention period, rather than effects on unobserved factors.

On the other hand, our results also provide sobering evidence against theories of skill building positing a very high self-productivity of early math skills, as the slope term observed for the post-test impact size was far below 1. This provides further evidence that non-experimental studies reporting links between skills over time likely overestimate the potential long-term effects of skill-focused interventions (see Bailey et al., 2018).

Additional Predictors of Follow-Up Impact

First, we found small associations between persistence conditional on post-test impacts and implementation quality. This may not seem surprising, as implementation quality is expected to improve student learning during the intervention period. Thus, the end-of-treatment impact measure will likely capture this influence. However, we found that blocks with the largest treatment-control contrast in time on mathematics generally showed less persistence than predicted based on their end-of-treatment impacts. Although this effect was not large, it might indicate that blocks that produced the largest treatment-induced gain in time on mathematics also include students in the control group who have received less math instruction and are thus likely to respond strongly to greater math instruction in kindergarten. Thus, for these children, kindergarten mathematics instruction likely acts as a substitute for preschool math instruction. This finding underscores the importance of considering what kinds of instruction students in the counterfactual receive—not just during but *after* the end of treatment.

In contrast to the predictions made by theories of dynamic complementarity (e.g., Johnson & Jackson, 2019), we found that blocks characterized by more learning in the post-intervention period across the treatment and control groups showed smaller follow-up treatment impacts. Disaggregating this "total" gains measure showed that learning gains by the control group in the post-intervention period were a strong negative predictor of follow-up impact. The

relatively weak prediction for post-intervention learning gains made by the treatment group was likely explained by the strong association between treatment gains and post-test impact magnitude, indicating that blocks with large treatment impacts were unlikely to produce high levels of treatment group learning following the intervention. Because of this strong association, the prediction for treatment group gains was minimal when accounting for post-test impact.

Some features of the post-intervention instructional environment may contribute to this negative correlation between post-test impact magnitude and treatment group learning. The idea that, in absolute terms, growth might slow following a period of quick learning, follows from empirical regularities, such as the "deceleration" of score growth on vertically scaled achievement tests (e.g., Student, 2022) and downward sloping learning curves (a point made related to fadeout by Campbell and Frey, 1970), along with the theoretical regularity that forgetting is paradoxically common following a period of rapid learning (Wixted, 2004). Indeed, a previous analysis of some of the sites included in this analysis found higher forgetting rates for students in the treatment group than for students in the control group in the year following the intervention (Kang et al., 2019).

Thus, we caution against a simplistic interpretation of this result as implying that an accelerated kindergarten curriculum would eliminate fadeout. This interpretation is challenged by our finding that environments with faster post-intervention gains showed weaker persistence, but also by other work showing that more advanced instructional content offerings in kindergarten do not explain differences in fadeout and persistence (Jenkins et al., 2018).

Although counterintuitive, we believe these findings hold important implications for our understanding of fadeout and persistence. First, they provide corroboration that post-intervention contexts are relevant for understanding fadeout dynamics. However, most recent work in this

area has theorized that early childhood intervention impacts will only persist if they are followed by higher-quality learning experiences (e.g., McCormick et al., 2022). Our results suggest, instead, that high-quality learning opportunities following an early intervention might lead to *more* fadeout if these learning opportunities are extended with equal likelihood to children in the treatment and control groups, which should be the case given children's randomization to conditions. Thus, any educational policy that will encourage learning by children in the control group, while socially desirable, might also lead to less persistent impacts for a given early intervention.

For educational policy, our results suggest that program developers should find settings where children are unlikely to receive similar learning opportunities to those offered by the intervention once the intervention ends. For example, targeting districts with low rates of subsequent growth (e.g., as estimated using administrative and NAEP data by Reardon, 2019) may plausibly generate more persistent impacts. Yet, when preschool programs are taken to scale, all children become "treated," and the key policy issue then involves promoting the highest learning rate possible for children during and after preschool. Our study is only applicable to the learning environments that were present in the kindergarten settings included in our data. We acknowledge the possibility that high-quality learning environments could be developed that would tailor instruction sufficiently well to students' individual learning trajectories in a way that maximizes the learning rate of each child, and ensures that early program impacts persist at higher rates. Of course, it is difficult to determine what the features of such "high-quality" environments might be. For example, some might consider teaching to be "high-quality" if it involves a suite of characteristics, such as a warm learning environment, positive teacher-student interactions, play-based activities, and child activity choice. However, the same instruction may

also have limited goals that address only material that treatment children have already learned (e.g., Engel et al., 2013). Common in the U.S., such teaching also might not include small group instruction or formative assessments, leading to a leveling of what is (or can be) learned (e.g., Black & William, 1998; Clements et al., 2023b).

Limitations

This analysis was intended to explore some of the processes hypothesized to underlie the persistence and fadeout of effective ECE interventions with the hope of improving our ability to forecast impacts in educational evaluation studies. However, for several reasons, our work does not provide a definitive test of any particular theory. The first limitation, as noted above, is that estimates of the associations between hypothesized moderators and kindergarten impacts were imprecise and are consistent with estimates much larger and much smaller than what we report. Unfortunately, analyses of block-level heterogeneity require very large numbers of blocks to obtain precise estimates (Weiss et al., 2017).

Second, the generalizability of these findings is limited. Although the children and settings under study are quite diverse within the U.S. context, the intervention under study is an ECE math intervention in all cases. Interventions targeting skills at different ages may show different patterns of persistence and fadeout. We hope that future work will use this analysis as a prototype for studying theories of persistence and fadeout as applied to different interventions.

Third, the estimates for moderators reported in this analysis are not causally identified and vary in the extent to which they are plausibly causal estimates (for review, see Rohrer & Arslan, 2021). On the one hand, control and treatment group learning in the post-intervention period are clearly causally related to persistence: A hypothetical intervention that affected one of these would affect persistence. On the other extreme, to our finding that larger impacts on non-

math skills negatively predicted follow-up math impacts, we are skeptical that a supplementary intervention that improved both children's math and non-math skills would show more fadeout than those included in the current study. Rather, the association between block-level impacts on non-math skills and fadeout more likely reflects something about the contexts in which the intervention positively impacted non-math skills. Future work that can identify exogenous variation in instructional features, for example, would be necessary to answer whether impacts on non-math skills would cause more fadeout or more persistence.

Finally, although our data were drawn from several scale-up cluster RCTs, each intervention study included design idiosyncrasies and limitations, including the use of inconsistent measures, that made it difficult to align data perfectly across the sites. These issues were reviewed above, and we found that results were generally robust to models that sequentially dropped observations from each respective site. Still, we are careful to label our study as "quasi-experimental," because we relied on child-level analyses within each block that likely depended heavily on pretest adjustments to make children comparable. Combined with the design issues reviewed above, including some departures from random assignment, our results should be replicated in other samples with rigorous experimental or quasi-experimental designs before drawing strong conclusions.

Conclusion

Our work suggests that post-test impacts are a strong predictor of one-year follow-up impacts across scale-up studies for an early childhood mathematics intervention. We found that impacts faded, on average, by about 60% of the initial post-test magnitude over one year, and that greater post-intervention learning in the control group additionally predicted more fadeout. Our work suggests that policies designed to improve post-intervention instructional quality may

lead to more fadeout if these policies also benefit the learning of children in the control group. Future work should seek to understand tradeoffs between boosting post-intervention learning and maximizing intervention impact by varying features of treatments, settings, and post-intervention redundancy.

Notes

[1] In San Diego, 10 classrooms were randomized instead of schools within the designated blocking groups. These classes are treated like "schools" in our analyses, and we do not include any further blocking variables at the school level for these classrooms.

[2] An analysis of end-of-preschool (i.e., endline) variation across the treatment sites was pre-registered (https://osf.io/dm697), with the plan of using a fixed intercept random coefficient model analysis. Initial results were presented at the 2020 Association for Public Policy Analysis and Management conference, and most hypothesized moderation effects were null. This analysis was largely exploratory in nature, and the pre-registration plan was simply a way to document our planned analyses before running models. A draft reporting these results is currently in progress. The current analyses of fadeout and persistence of follow-up effects were not pre-registered and are also considered exploratory.

[3] This exclusion restriction essentially dropped blocks that had no variation in our key treatment assignment due to schools dropping out of the study or switching assignments (an issue that was limited in scope, but we discuss it in more detail below).

[4] NYC used an abbreviated version of COEMET that contained a reduced number of Likert scale items (see Morris et al., 2016 for more information). To produce a measure of instructional quality that was consistent across study sites, we restricted all COEMET instructional quality ratings to the common items available in NYC and the other sites.

[5] This estimate does not directly correspond to the overall block-level average shown in Table 2, but was generated from a meta-regression that weights estimates based on precision. The average post-test and follow-up impacts without pre-test adjustments were 0.26 and 0.14, respectively.

[6] COEMET-related moderators (impacts on math instructional time, number of math activities, and overall COEMET quality score) were examined using data from only 40 blocks. One block was excluded from these exploratory analyses due to missing data on math instructional quality items, which impacted within-block treatment-control variation.

References

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological Bulletin, 145*(12), 1103–1127. https://doi.org/10.1037/bul0000212

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020a). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest, 21*(2), 55–97. https://doi.org/10.1177/1529100620915848

Bailey, D. H., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist, 73*(1), 81–94. https://doi.org/10.1037/amp0000146

Bailey, D. H., Jenkins, J. M., & Alvarez-Vargas, D. (2020b). Complementarities between early educational intervention and later educational quality? A systematic review of the sustaining environments hypothesis. *Developmental Review, 56*, 100910. https://doi.org/10.1016/j.dr.2020. 100910

Bailey, D. H., Nguyen, T., Jenkins, J. M., Domina, T., Clements, D. H., & Sarama, J. S. (2016). Fadeout in an early mathematics intervention: Constraining content or pre-existing differences? *Developmental Psychology, 52,* 1457-1469.

Bierman, K. L., Nix, R. L., Heinrichs, B. S., Domitrovich, C. E., Gest, S. D., Welsh, J. A., & Gill, S. (2014). Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *Child Development, 85*(1), 140-159.

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start* (National Bureau of Economic Research Working Paper Series, No. 20434). https://doi. org/10.3386/w20434

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–76.

Botvin, C. M., Jenkins, J. M., Carr, R. C., Dodge, K. A., Clements, D. H., Sarama, J., & Watts, T. W. (2024). Can peers help sustain the positive effects of an early childhood mathematics intervention?. *Early Childhood Research Quarterly*, *67*, 159-169.

Braithwaite, D. W., & Siegler, R. S. (2023). A unified model of arithmetic with whole numbers, fractions, and decimals. *Psychological Review*.

Campbell, D. T., & Frey, P. W. (1970). The implications of learning theory for the fade-out of gains from compensatory education. Compensatory Education: A National Debate, 3, 455–463.

Carolan, M. (2014, September 16). *"Fadeout" in early childhood: Does the hype match the research?* National Institute for Early Education Research (NIEER). https://nieer.org/research-library/fadeout-early-childhood

Carr, R. C., Jenkins, J. M., Watts, T. W., Peisner-Feinberg, E. S., & Dodge, K. A. (2024). Investigating if high-quality kindergarten teachers sustain the pre-K boost to children's emergent literacy skill development in North Carolina. *Child Development.*

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics, 126*(4), 1593-1660.

Claessens, A., Engel, M., & Curran, F. C. (2014). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal, 51*(2), 403-434.

Clark, C. A., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental psychology*, *46*(5), 1176.

Clements, D. H., Lizcano, R., & Sarama, J. (2023b). Research and pedagogies for early math. *Education Sciences, 13*(839). https://doi.org/10.3390/educsci13080839

Clements, D. H., Sarama, J., Layzer, C., & Unlu, F. (2023a). Implementation of a scale-up model in early childhood: Long-term impacts on mathematics achievement. *Journal for Research in Mathematics Education*, *54*(1), 64-88. https://doi.org/10.5951/jresematheduc-2020-0245

Clements, D. H., Sarama, J., Layzer, C., Unlu, F., & Fesler, L. (2020). Effects on mathematics and executive function of a mathematics and play intervention versus mathematics alone. *Journal for Research in Mathematics Education*, *51*(3), 301-333.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based early maths assessment. *Educational Psychology, 28*(4), 457–482.

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.

Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). *TEAM—Tools for early assessment in mathematics.* McGraw-Hill Education.

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

Cunha, F., & Heckman, J. J. (2007). The technology of skill formation. *American Economic Review, 97*(2), 31–47.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K. A., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446. https://doi. org/10.1037/0012-1649.43.6.1428

Duncan, G. J., & Sojourner, A. J. (2013). Can intensive early childhood intervention programs eliminate income-based cognitive and achievement gaps? *Journal of Human Resources*, *48*(4), 945-968.

Durkin, K., Lipsey, M. W., Farran, D. C., & Wiesen, S. E. (2022). Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade. *Developmental Psychology, 0*, 470–484. https://doi.org/10.1037/dev0001301

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2016). Early childhood education. In *Economics of means-tested transfer programs in the United States, volume 2* (pp. 235-297). University of Chicago Press.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35*(2), 157–178. https://doi.org/10.3102/ 0162373712461850

Gray-Lobe, G., Pathak, P. A., & Walters, C. R. (2023). The long-term effects of universal preschool in Boston. *The Quarterly Journal of Economics*, *138*(1), 363-411.

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., & Watts, T. W. (2024). Fadeout and persistence of intervention impacts on social–emotional and cognitive skills in children and adolescents: A meta-analytic review of randomized controlled trials. *Psychological Bulletin*, *150*(10), 1207.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives*, *2*(3), 172-177.

Hofer, K. G., Lipsey, M. W., Dong, N., & Farran, D. C. (2013a). Results of the Early Math Project--Scale-Up Cross-Site Results. Working Paper. *Peabody Research Institute*.

Hofer, K. G., Farran, D. C., & Cummings, T. P. (2013b). Preschool children's math-related behaviors mediate curriculum effects on math achievement gains. *Early Childhood Research Quarterly, 28*(3), 487-495.

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. J. (2018). Do high-quality kindergarten and first-grade classrooms mitigate preschool fadeout? *Journal of Research on Educational Effectiveness, 11*(3), 339–374.

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy*, *11*(4), 310-349.

Kang, C. Y., Duncan, G. J., Clements, D. H., Sarama, J., & Bailey, D. H. (2019). The roles of transfer of learning and forgetting in the persistence and fadeout of early childhood mathematics interventions. *Journal of Educational Psychology*, *111*(4), 590.

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics*, *131*(4), 1795-1848.

Li, W., Duncan, G. J., Magnuson, K., Schindler, H. S., Yoshikawa, H., & Leak, J. (2020). *Timing in early childhood education: How cognitive and achievement program impacts vary by starting age, program duration, and time since the end of the program* (EdWorkingPaper: 20-201). Annenberg Institute at Brown University. https://doi.org/10.26300/5tvg-nt21

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly, 45*, 155–176. https://doi.org/10.1016/j.ecresq.2018.03.005

McCormick, M. P., Pralica, M., Weiland, C., Hsueh, J., Moffett, L., Guerrero-Rosada, P., ... & Sachs, J. (2022). Does kindergarten instruction matter for sustaining the prekindergarten (PreK) boost? Evidence from individual-and classroom-level survey and observational data. *Developmental Psychology*, *58*(7), 1298.

Morris, P. A., Mattera, S. K., & Maier, M. F. (2016). Making Pre-K Count: Improving Math Instruction in New York City. *MDRC*.

Najarian, M., Snow, K., Lennon, J., Kinsey, S., & Mulligan, G. (2007). Early childhood longitudinal study, birth cohort (ECLS-B).

Pearman, A., Springer, M., Lipsey, M., Lachowicz, M., Swain, W., & Farran, D. (2020). Teachers, schools, and pre-k effect persistence: An examination of the sustaining

environment hypothesis. *Journal of Research in Educational Effectiveness.* https://doi.org/10.1080/19345747.2020.1749740

Reardon, S. F. (2019). Educational opportunity in early and middle childhood: Using full population administrative data to study variation by place and age. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *5*(2), 40-68.

Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007368.

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist, 57*(1), 31-54.

Student, S. R. (2022). Vertical scales, deceleration, and empirical benchmarks for growth. *Educational Researcher*, *51*(8), 536-543.

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

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352–360.

Watts, T. W., Jenkins, J. M., Dodge, K. A., Carr, R. C., Sauval, M., Bai, Y., Escueta, M., Duer, J., Ladd, H., Muschkin, C., Peisner-Feinberg, E., & Ananat, E. (2023). Understanding heterogeneity in the impact of public preschool programs. *Monographs of the Society for Research in Child Development*, *87*(4).

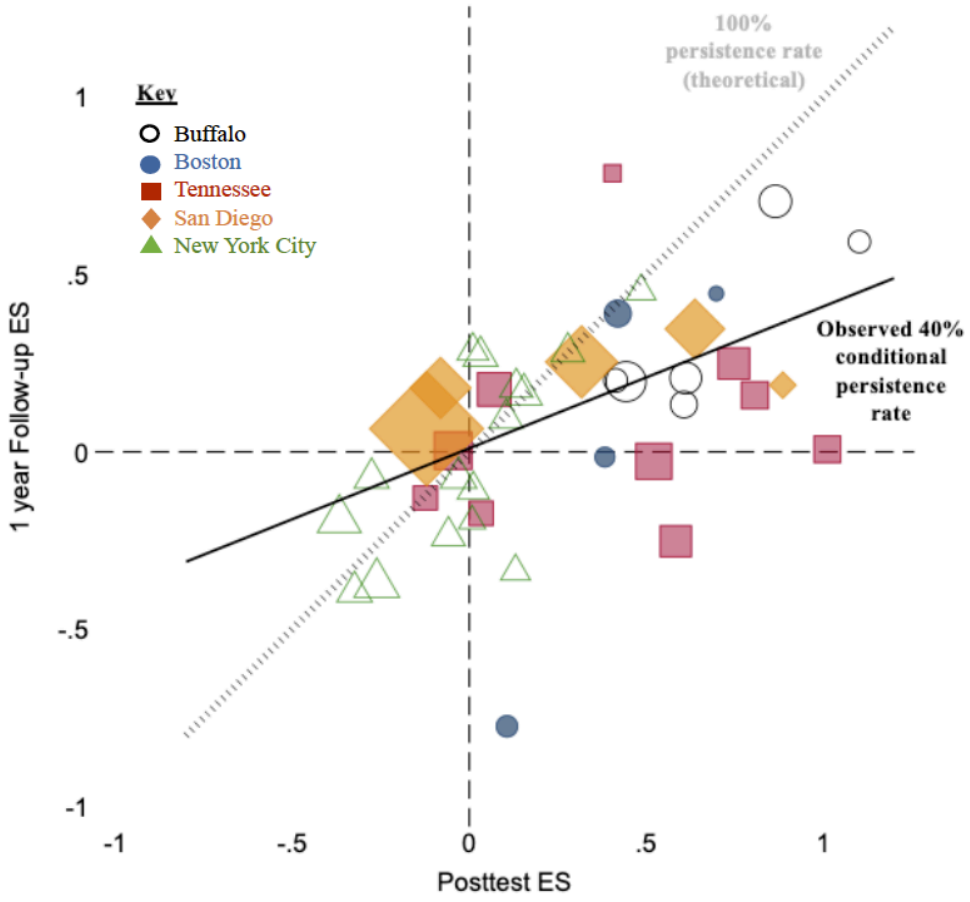
Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843-876.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.*, *55*(1), 235-269.

Woodcock, Richard W., Kevin S. McGrew, and Nancy Mather. 2001. Woodcock-Johnson III Tests of Achievement. Riverside Publishing.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology*, *50*(12), 2572.

Figure 1
Empirical association between post-test and follow-up impacts across sites



Note. This figure directly corresponds to the model results shown in Column 2 of Table 3. Each marker represents a different block, with the marker size weighted by 1/se2 of the follow-up impact estimate. The gray dashed line represents the regression line that would be observed with 100% conditional persistence between posttest and follow-up. The black line represents the observed relation (i.e., B = 0.40, SE = .13).

Table 1

Child-Level Descriptives Statistics Across Study Sites

	Buffalo	**Boston**	**Tennessee**	**San Diego**	**NYC**	**Full Sample**
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Female	0.51	0.49	0.55	0.54	0.52	0.53
Black- Non Hispanic	0.56	0.40	0.77	0.05	0.35	0.49
White- Non Hispanic	0.26	0.11	0.10	0.24	0.03	0.11
Ethnicity- Other	0.03	0.09	0.04	0.23	0.04	0.05
Hispanic	0.15	0.40	0.09	0.48	0.58	0.35
Limited Eng Prof.	0.06	0.35	0.10	0.44	0.10	0.13
Age at Pre-K Entry (years)	4.25 (0.31)	4.64 (0.31)	4.45 (0.31)	4.39 (0.40)	4.30 (0.29)	4.37 (0.34)
Age at Pre-K Post (years)	4.92 (0.30)	5.31 (0.32)	5.05 (0.31)	4.99 (0.33)	4.83 (0.30)	4.95 (0.33)
Observations (Children)	521	174	771	699	1216	3381

Note. Means and standard deviations (in parentheses) are presented separately for each study site. Descriptive information for the entire sample (n = 3,381) is provided in the final column. Note that the TN site had substantial missing data on child demographic characteristics (discussed further in the text; see supplementary file).

Table 2

Block Level Descriptive Statistics Across Study Sites

	All Sites		**Buffalo**		**Boston**		**Tennessee**		**San Diego**		**NYC**	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Post-Test Impacts (Adj. for Pre-Test)	0.28	0.38	0.67	0.26	0.41	0.25	0.41	0.41	0.32	0.43	0.00	0.22
Follow-Up Impacts (Adj. for Pre-Test)	0.09	0.31	0.33	0.24	0.01	0.63	0.08	0.31	0.21	0.11	-0.01	0.27
Post-Test Impacts (No Pre-Test Adj.)	0.30	0.37	0.55	0.42	0.21	0.68	0.51	0.28	0.21	0.30	0.12	0.25
Follow-Up Impacts (No Pre-Test Adj.)	0.11	0.32	0.23	0.37	-0.09	0.88	0.09	0.17	0.12	0.30	0.12	0.23
Total Gains Between Prek-Post and K	1.29	0.19	1.26	0.17	1.32	0.16	1.26	0.09	0.98	0.22	1.40	0.14
Control Group Gains Between Prek-Post and K	1.41	0.23	1.45	0.21	1.49	0.13	1.49	0.15	1.01	0.32	1.44	0.14
Treatment Group Gains Between Prek-Post and K	1.17	0.24	1.07	0.18	1.14	0.21	1.04	0.16	0.94	0.20	1.36	0.18
Total Post-Test Level	0.19	0.31	-0.04	0.36	0.03	0.24	0.00	0.15	0.22	0.63	0.02	0.21
Control Group Post-Test Level	0.05	0.34	-0.31	0.45	-0.08	0.37	-0.25	0.24	0.12	0.67	-0.04	0.22
Treatment Group Post-Test Level	0.34	0.37	0.24	0.37	0.13	0.46	0.25	0.16	0.32	0.61	0.08	0.27
Non-Math Cognitive Impacts (Prek)	0.02	0.33	-0.05	0.25	-0.21	0.12	0.06	0.42	0.05	0.33	0.09	0.30
Impact on Overall COEMET Quality Score (Prek)	0.52	1.22	0.05	1.07	0.22	1.00	0.92	1.08	0.86	0.97	0.10	0.65
Impact on Math Instructional Time (Prek)	1.71	2.15	0.92	0.70	1.68	1.32	1.36	0.76	0.42	0.23	0.72	0.46
Impact on Number of Math Activities (Prek)	1.08	1.01	0.84	1.16	0.76	0.90	1.10	0.67	0.63	0.58	0.76	0.71
Limited English Proficiency	0.15	0.16	0.05	0.09	0.30	0.15	0.11	0.10	0.42	0.23	0.10	0.07
Black-Non Hispanic	0.46	0.30	0.59	0.29	0.42	0.11	0.74	0.23	0.06	0.03	0.36	0.24
Observations	41		6		4		10		5		16	

Note. Means and standard deviations are presented for each study site. All values are presented at the block level (n = 41). Post-test and follow-up impacts were derived from child-level analyses in which our standardized composites of math achievement were regressed on an indicator for treatment status, child demographic controls, and the pre-k entry pre-test score. The post-test and follow-up impacts without pre-test adjustment were derived with no controls included in the models. The gains estimates were taken by calculating the non-standardized difference between the kindergarten math score and end-of-pre-k math score. This difference was then divided by the end-of-pre-k math score standard deviation in the control group to derive a common scale across sites. The post-test-level variables provide block-level estimates of the standardized post-test composite across sites. The non-math cognitive impacts, impact on overall COEMET quality score, impact on math instructional time, and impact on number of math activities variables were derived from child-level regression models that regressed each respective score on the treatment indicator, demographic controls, and pretest.

Table 3

Block-Level Predictors of Follow-Up Impact

	(1)	(2)	(3)	(4)	(5)
Post-Test Impacts (Adj for Pre-Test)	0.40**	0.39**	0.44***	0.41**	0.36+
	(0.13)	(0.12)	(0.12)	(0.16)	(0.21)
Total Post-Test Level		-0.16			
		(0.14)			
Total Gains Between Pre-K Post and K		-0.45+			
		(0.24)			
Control Group Post-Test Level			-0.22+		-0.61*
			(0.13)		(0.26)
Control Group Gains Between Pre-K Post and K			-0.53**		-0.91***
			(0.20)		(0.27)
Treatment Group Post-Test Level				-0.01	0.50+
				(0.15)	(0.28)
Treatment Group Gains Between Pre-K Post and K				0.02	0.67*
				(0.30)	(0.34)
Constant	0.01	0.60+	0.71**	-0.02	0.34
	(0.06)	(0.32)	(0.27)	(0.39)	(0.34)
N (City / Study Blocks)	5 / 41	5 / 41	5 / 41	5 / 41	5 / 41
τ (site)	0.06	0.00	0.00	0.08	0.00

Note. + p < 0.1 *p < 0.05. **p < 0.01. ***p < 0.001. Standard errors are provided in parentheses. All models are run at the block level (n=41) and control for children's pre-test assessment. Average effect sizes were estimated using a random effects meta-analytic model that included a random effect for study site and weights. The tau represents between study variation in effects (interpreted in standard deviation units).

Table 4

Block-Level Exploratory Predictors of Follow-Up Impact

	(1)	(2)	(3)	(4)	(5)	(6)
Post-Test Impacts (Adj for Pre-test)	0.47***	0.42**	0.47***	0.43**	0.44***	0.46***
	(0.14)	(0.14)	(0.12)	(0.14)	(0.12)	(0.13)
Non-Math Cognitive Impacts (Pre-K; Adj. for Pre-Test)	-0.20					
	(0.16)					
Impact on Overall COEMET Quality Score (Pre-K)		-0.01				
		(0.04)				
Impact on Math Instructional Time (Pre-K)			-0.06*			
			(0.03)			
Impact on Number of Math Activities (Pre-K)				-0.02		
				(0.05)		
Limited English Proficiency (10%)					0.03	
					(0.02)	
Black-Non Hispanic (10%)						-0.02
						(0.01)
Constant	-0.01	0.01	0.07	0.02	-0.06	0.07
	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
N (Site / Study Blocks)	5 / 41	5 / 40	5 / 40	5 / 40	5 / 41	5 / 41
τ (Site)	0.08	0.07	0.00	0.06	0.00	0.00

Note. + p < 0.1 *p < 0.05. **p < 0.01. ***p < 0.001. Standard errors are provided in parentheses. All models are run at the block level (n=41) and control for children's pre-test assessment. Average effect sizes were estimated using a random effects meta-analytic model that included a random effect for study site and meta-analytic weights. The tau represents between study variation in effects (interpreted in standard deviation units). Both Limited English Proficiency and Black Non-Hispanic variables were rescaled to 10 percentage point units.