The Sequential Scale-Up of an Evidence-Based Intervention: A Case Study

Jaime Thomas

Mathematica Policy Research

Thomas D. Cook

Northwestern University and George Washington University

Alice Klein and Prentice Starkey

WestEd

Lydia DeFlorio

University of Nevada, Reno

June 2018

Corresponding Author: Jaime Thomas, Mathematica Policy Research, 505 14th Street, Suite 800, Oakland, CA 94612; Phone: 510-830-3717; Fax: 510-830-3701; Email: jthomas@mathematica-mpr.com.

Acknowledgements: The authors would like to acknowledge Stacey Koo and Laura Gluck from WestEd for invaluable research assistance.

The research reported here was supported by Grant U411B120053 from the Office of Innovation and Improvement (OII), U.S. Department of Education to WestEd. The opinions expressed are those of the authors and do not necessarily represent views of OII or the U.S. Department of Education. This work was also supported by the following: Charles and Lynn Schusterman Family Foundation, David and Lucile Packard Foundation, East Bay Community Foundation, Edmodo, Grand Canyon University, IBM, Heising-Simons Foundation, S.D. Bechtel Jr. Foundation, Stuart Foundation, and University of Oregon Foundation.

Citation for published version of this article:
Thomas, J., Cook, T. D., Klein, A., Starkey, P., & DeFlorio, L. (2018). The Sequential Scale-Up of an Evidence-Based Intervention: A Case Study. *Evaluation Review*, 0193841X18786818.

ABSTRACT

Policymakers face dilemmas when choosing a policy, program, or practice to implement. Researchers in education, public health, and other fields have proposed a sequential approach to identifying interventions worthy of broader adoption, involving pilot, efficacy, effectiveness, and scale-up studies.

In this paper, we examine a scale-up of an early math intervention to the state level, using a cluster randomized controlled trial. The intervention, *Pre-K Mathematics*, has produced robust positive effects on children's math ability in prior pilot, efficacy, and effectiveness studies. In the current study, we ask if it remains effective at a larger scale in a heterogeneous collection of pre-K programs that plausibly represent all low-income families with a child of pre-K age who live in California. We find that *Pre-K Mathematics* remains effective at the state level, with positive and statistically significant effects (effect size = 0.30, $p < 0.01$).

In addition, we develop a framework of the dimensions of scale-up to explain why effect sizes might decrease as scale increases. Using this framework, we compare the causal estimates from the present study to those from earlier, smaller studies. Consistent with our framework, we find that effect sizes have decreased over time.

We conclude with a discussion of the implications of our study for how we think about the external validity of causal relationships.

1. INTRODUCTION

Policymakers at the local, state, or national level face dilemmas when choosing a policy, program, or practice to implement. Researchers in education, public health, and other fields have proposed a sequential approach to identifying interventions worthy of being recommended for broad adoption. This sequential approach involves classes of inquiry that increase in scope and vary in purpose.

Pilot studies examine the feasibility of an intervention that has, along with its evaluation, been implemented under highly controlled conditions (Leon, Davis, & Kraemer, 2011). They are designed to answer the question: “Is the theory undergirding the intervention effective?” *Efficacy research* assesses the extent to which a successfully piloted intervention produces the desired outcomes under less controlled conditions that are still local enough for the developer to be actively involved in implementing and evaluating the program. The research question here is: “Can the intervention work outside of the laboratory or a few carefully selected sites under conditions that enable high quality implementation?”

Effectiveness studies are less controlled, and test whether an intervention is still effective under conditions that approximate the real world of intended application even though the scale is still smaller than a policymaker’s remit, with less developer involvement in implementation and evaluation (Earle et al., 2013). The question here is: “Does the program work under conditions that approximate those under which the intervention would be delivered on a broader scale?”

The fourth class of inquiry consists of *scale-up studies.* These studies examine whether a successfully evaluated intervention continues to be effective when all the conditions that would be in place if the intervention were official policy are approximated in the study (Flay et al., 2005; Gottfredson et al., 2015). Scale-up studies can be characterized in two ways. Relative scale-up describes a study that is merely larger than earlier studies, and so some uncertainty remains about identifying the target population—because no rules are evident for inferring the target population from the sampling specifics. Absolute scale-up has the goal of identifying program effectiveness in a specific population like a nation, state, or local government, and also entails a valid justification for the necessary extrapolation from the obtained sample to the intended population.

The preferred social science rule for extrapolation involves drawing a probability sample from a clearly designated population, thereby ensuring that the sample represents its target population within known limits of sampling error. Although random selection is a method widely used to describe human populations, it is rarely used to generalize knowledge about causal connections. More common are attempts to ground claims about causal generalization in large, heterogeneous, and purposively selected samples of sites and individuals from within the target population, choosing them so as to vary all the factors that are believed to condition the effect or that study stakeholders believe are important. In studies where population attributes are available in detail, it is then possible to weight the study characteristics that were assessed to better approximate the population, at least on measures common to the population description and the available measures of the sample.

In this paper, we examine a case of absolute scale-up of an early math intervention to the state level. The intervention, *Pre-K Mathematics*, has produced robust positive effects on children's math ability in prior pilot, efficacy, and effectiveness studies. In the current study, we ask: Does it remain effective when implemented at about double the prior scale in a more heterogeneous collection of purposively chosen public pre-kindergarten (pre-K) and Head Start programs that plausibly represent all low-income families with a child of pre-K age who live in California, the nation's largest state?

In addition, we develop a framework of the dimensions of scale-up to explain why effect sizes might change as scale increases. Using this framework, we compare the causal estimates from the present study to those from earlier, smaller studies. Taken together, this set of studies represents an even larger sample of student respondents, treatment providers, pre-K settings, time periods, and different ways of operationalizing both the *Pre-K Mathematics* treatment and the math outcome measure. As a result, the heterogeneity across studies is even greater than the heterogeneity within the latest and most heterogeneous study in the programmatic sequence we explore. Examining a scale-up study within the program of studies it is part of permits an even broader examination of the external validity of any named intervention.

The paper has six sections. First, we describe our framework of the dimensions of scale-up. Second, we describe the *Pre-K Mathematics* program and past evaluations of it. Third, we describe the methods we used to assess the effects of *Pre-K Mathematics* in the current study and to compare these effects with those found in earlier studies. Fourth, we test whether the version of *Pre-K Mathematics* evaluated in this study raises average math achievement at the statewide scale, and whether the program effects we obtain are robust across a wide range of student and site attributes. Fifth, we describe how effect sizes differ between the present study and earlier studies, in particular asking whether effect sizes tend to diminish as study samples get larger and more heterogeneous. Finally, we discuss the implications of our findings, both for scale-up studies in particular and for external validity and causal generalization writ large.

1.1. Relevant Theory

In this section, we provide a framework of the dimensions of scale-up and how each of these dimensions can influence effect sizes. In this framework, scale-up influences the number and heterogeneity of settings and study participants, the quality of the program content and its delivery, changes in the counterfactual or "business-as-usual" condition, variations in the outcome measure and measurement quality, and the quality of the evaluation's design and execution. Figure 1 depicts the framework and the hypothesized influences of scale-up on effect sizes.

[Figure 1 about here]

1.1.1. Number and Heterogeneity of Settings and Study Participants

Definitions of scale-up research tend to focus on the size of study samples. McDonald, Keesler, Kauffman, & Schneider (2006) emphasize the increased number of settings and participants that typically accompany scale-up efforts, but they do not specify how this increase might influence effect sizes. The Common Guidelines for Education Research and Development also describe scale-up research as an increase in study size, but the guidelines also link the increase in size to sample heterogeneity and researcher independence when emphasizing "effectiveness in a wide range of populations, contexts, and circumstances, without substantial developer involvement in implementation or evaluation" (Earle et al., 2013, p. 9).

As the number of settings and the size of study populations rise due to scale-up, the variation in those settings and populations is likely to increase as well. Although studies of academic achievement usually reveal more variation within sites than between them (Tipton, Hallberg, Hedges, & Chan, 2016), a large national study of 4-year-olds will reflect a greater range of student abilities, race/ethnicity profiles, home circumstances, prior family exposure to pre-K services, and the like compared to what any small, local study can provide. The same is true for the range of parent, service provider, and site attributes. It is such sampling heterogeneity that allows scale-up studies to probe impacts across a wider range of characteristics than earlier studies could.

It is not entirely clear how sample differences between earlier and later studies influence effect sizes. Such compositional differences will only lead to differences in treatment effects if the earlier and later study samples differ in subgroup characteristics that influence effect sizes. When they do, whether compositional differences increase or decrease effect size depends on whether the earlier or later study sample has attributes related to larger or smaller effects. In this connection, it is particularly important to test whether any study results in negative effects—as inferred from statistically significant effects with the opposite causal sign. Such effects are "iatrogenic," and their ethical and political implications need serious consideration.

Compositional differences between studies matter technically as well as substantively. Researchers use effect sizes to compare causal estimates within and between studies, and Formula 1 (below) describes the calculation of effect sizes. The numerator indexes the size of the mean post-test difference between the treatment and control groups in the original study metric. The denominator is an estimate of the standard deviation—either the control group standard deviation, or the pooled control and treatment group standard deviations when they do not differ. The effect size is the ratio of these two values, and it transforms the estimate from its original metric into a standard deviation metric that can then be compared across outcomes within and between studies.

$$(1)\quad Effect\ size = \frac{Treatment\ group\ mean - Compariso\ \ \ group\ mean}{Standard\ deviation}$$

Now imagine an intervention that is more effective with children who initially score lower on an achievement test, and for which the proportion of lower achievers is higher in the scale-up study than in prior studies. The numerator of Formula 1 will be larger in the scale-up study than in earlier studies and will result in larger effect sizes. Conversely, if the intervention were more effective for initially high-scoring children, then the scale-up's numerator would be lower, and its effect size would be smaller.

Formula 1 also requires standard deviations and, if the larger scale-up sample is—as one would expect—more heterogeneous than samples from prior studies, the scale-up study will generate a larger standard deviation. For a given mean difference in the numerator, a larger standard deviation in the denominator would reduce the effect size. To make this clearer, imagine two evaluations of the same intervention. The first is conducted in a single site with children who are homogeneous in their socioeconomic background, their pre-intervention achievement, and other characteristics that affect achievement. The second is conducted with a statewide sample of children. All things being equal, the standard deviation will tend to be smaller in the single-site study because of the homogeneity of its sample, and the effect sizes would consequently be larger.

In actual research practice, it is difficult to predict how sample heterogeneity will influence effect sizes—will the effect size increase because participants who benefit more from the intervention are now part of the sample in the larger study? Or will it decrease because the sample now includes participants who benefit less from the intervention, or because a larger standard deviation in the denominator swamps an increase in the numerator? The ambivalence of the relationship between sample size and heterogeneity and effect sizes is reflected in Figure 1, where a two-headed arrow (instead of a single-headed one) is used to illustrate how the added heterogeneity of many scale-up studies might influence effect sizes.

1.1.2. Quality of Program Content and Delivery

Glennan, Bodilly, Galegher, and Kerr (2004) characterize scale-up as a "non-sequential process of interaction, feedback, and adaptation among groups of actors" (p. 27). This definition implies that larger scales may entail variation in program content and delivery as well as in sample size and heterogeneity.

Although positive changes in content and delivery may occur when moving from the smaller to the bigger scale, there are many more reasons to believe such changes will be negative and reduce effect sizes.

On the positive side, many developers and implementers refine their programs to achieve ever more consistent and reliable results over time (McDonald et al., 2006), leading to later implementations that are subtly different from earlier ones, even when the program name does not change. The hope is that such program changes will reflect lessons gained from experience and so increase the numerator in Formula 1.

But scale-up can also have a negative effect on implementation quality, principally by diminishing the developer's control and decreasing implementation fidelity. As the number of sites and treatment providers increases, developers will tend to be less involved in implementation and provide less support on things like training treatment providers and monitoring program activities (Earle et al., 2013). The consequences of this for Formula 1 are likely to be both a lower numerator and a larger denominator, and lower effect sizes as a result.

But budgets must also be taken into account. If the per-site and per-person budgets are constant across the earlier and later studies, or even higher in the latter, then the training, implementation, and monitoring disadvantages of having more sites could be avoided. But it seems naïve to assume research budgets that automatically increase to reflect the larger scaled-up samples, for this belies an important rationale for scaling up—that budgets reflect the lower cost patterns likely to hold when an intervention is standard practice and not a promising demonstration.

Our judgment is that, when it comes to the effect sizes that a study achieves, improved program design will be less consequential than diminished implementation quality. This belief rests on the assumption that a higher number and greater dispersal of scale-up sites will result in less intense and less productive relationships with program developers—for example, instead of the program developers training the program deliverers themselves, they would use indirect approaches, such as train-the-trainer or online methods. The result of such compromises might be that a larger proportion of local sites fail to adopt some program components, or that they inappropriately adapt others. In either case, the scaled-up intervention should be weakened when compared to smaller-scale implementations over which program developers can “hover” more readily. Figure 1 depicts the program implementation changes from scale-up as a single-headed arrow pointing downward to indicate smaller effects at a larger scale.

1.1.3. Changes in the Counterfactual Condition

Scale-up studies take place after results from a mix of smaller studies have been so consistently positive that they justify a larger study. This later study by definition takes place in a different time period, which is characterized by its own dynamics that might affect impact estimates. Indeed, many topics worthy of testing on a larger and more expensive scale may have to show not only prior evidence of effectiveness, but also evidence that they tap into types of social change that command current excitement and acceptance among the public and relevant decision makers.

Are scale-up studies more likely to happen when a policy agenda has crystallized around the importance and definition of both the problem and a broad class of acceptable solutions to it? If so, the intervention itself will not seem as novel as it did before, and—more important for our purposes—the comparison group used in the scale-up will reflect some part of this new policy agenda and thus perhaps demonstrate increases in performance over time. If the counterfactual condition that most experiments and quasi-experiments require has come to incorporate more elements that overlap with the content of the intervention, then it will still be a “business as usual” counterfactual, but one whose performance is different from what used to define business as usual. The results will be a higher hurdle for the treatment group to overcome in the scale-up study, and a corresponding reduction in the obtained effect size.

In this connection, consider the field of early childhood research. More and more children are attending pre-K, pre-K instruction has become more professionalized, and instructional content has become more aligned with elementary school standards. Indeed, the California Preschool Instructional Networks were developed during the period of this study, and the Common Core State Standards for Mathematics have focused on getting preschool children ready to meet the standards they will face in kindergarten (Lewis Presser, Clements, Ginsburg, & Ertle, 2015). It is almost certain, therefore, that the content of the business-as-usual counterfactual has shifted, and that pre-K students in the control group will now perform better. Such contextual changes will decrease the numerator in Formula 1 and so reduce effect sizes. Thus, Figure 1 includes a downward-facing arrow to reflect how secular changes in the counterfactual condition are hypothesized to reduce effect sizes.

1.1.4. Variations in the Quality of Outcome Measurement

In smaller studies, program developers and researchers can have considerable control over how measurement takes place. They can ensure that the conditions and timing of testing reflect the research plan, and that attrition from the testing regimen is minimal. At a larger scale, the chances of having consistently high quality measurement are lower. Measurement error will tend to be greater, the denominator in Formula 1 will increase, and smaller effect sizes should result.

However, two factors complicate fulfillment of this expectation. The first is that outcome measures can change over time. Program developers or researchers evaluating an intervention can choose to use different outcome measures—for example, they could move from a developer-created measure to a nationally normed test. Even if the same outcome measure is used, the measure itself can evolve over time, whether or not the abstract construct the measure is meant to represent does. Then, the later scale-up will likely involve at least a partially novel outcome measure. If the new measure is better aligned to the intervention content, it will increase the numerator in Formula 1; or if it has superior psychometric properties, it will decrease the denominator. In either case, the effect size would increase.

In contrast, moving from a developer-designed measure to a national- or state-normed test might negatively affect alignment (and, all else equal, the effect size), because no one knows the specifics of a program better than its developers do. It is hard to know how to weight these countervailing forces, though most math achievement tests are well constructed so that the marginal gains in their psychometric quality would seem modest in comparison with the potential loss in content alignment.

The second complicating factor is the research budget. In scale-up studies, the total budget will typically be higher than it is in smaller studies. However, the per-site or per-respondent budgets are more important for measurement quality, and they need not increase. Indeed, they are likely not to, because, as noted, one rationale for scaling up is to see what happens when the costs are those of an operative program instead of those motivated by the need for faithful implementation and high quality measurement as in smaller studies.

All other things being equal, we judge that the larger scale is likely to lead to less control over the measurement process and to lower reliability when compared to the earlier and more controlled studies in a programmatic sequence. Anecdotal evidence indicates that a portion of the outcome testing in some scale-up studies has taken place in unusual (and potentially more distracting) settings like homes, libraries, coffee houses, and cars in order to keep children who could not be tested in school within the study's measurement framework. Such practices will tend to inflate the denominator in Formula 1 and thus attenuate effect sizes. Hence the downward-facing arrow in Figure 1, which indicates lower effect sizes in later studies due to lower measurement quality.

1.1.5. Quality of the Evaluation Design and Its Execution

Most interventions selected for scale-up will already have evidence of their efficacy or effectiveness from high quality studies, including randomized controlled trials (RCTs). Although the quality of design plans might not differ much between earlier and later studies, control over the implementation of the evaluation—that is, the execution of the planned study design—will tend to be lower as the size and heterogeneity of a scale-up study increases. The net result is more slippage between the planned and achieved design, and a correspondingly lower likelihood of finding impacts as large as earlier ones found in smaller and more homogeneous studies, where there was greater control over implementing both the program and its evaluation. For example, a smaller experiment will likely entail less attrition, both overall and differential, and thus better pre-test balance in the post-attrition analytic sample, resulting in less biased impact estimates. The downward-facing arrow in Figure 1 suggests that in scale-up studies, the planned study will not be executed with the same faithfulness to its design, and hence will have smaller effect sizes.

1.1.6. Overall Implications for Effect Sizes

Figure 1 summarizes the expected changes in effect sizes as scale increases—first for each unique dimension of scale-up, then across all dimensions. Upward-facing arrows indicate larger effect sizes at larger scale, double-headed arrows indicate no clear prediction of direction, and downward-facing arrows suggest smaller effects. Weighting the dimensions equally, the final column indicates our belief that scaling up is more likely to decrease effect sizes than to increase them.

This is not a formal hypothesis, however, for an increase attributable to one factor might lead to an overall increase that is larger than the cumulative decreases attributable to the other factors. Moreover, the influence of each force compelling the study toward finding smaller effects can be mitigated by more care, more resources, and more foresight—not to mention more luck.

2. THE *PRE-K MATHEMATICS* PROGRAM AND PAST EVALUATIONS OF IT

In this section, we describe the intervention under study, *Pre-K Mathematics*, and situate the current study in the context of past evaluations of *Pre-K Mathematics*.

2.1. The Intervention

Pre-K Mathematics (Klein & Starkey, 2002; 2004) is a multicomponent supplementary math curriculum for pre-kindergarten children. Past studies at different scales have found unequivocal evidence of its effectiveness, and the What Works Clearinghouse (2013) currently rates it as having statistically significant and positive effects on math achievement. *Pre-K Mathematics* focuses on the pre-K classroom and home learning environments of young children, especially those from families experiencing economic hardship. Its activities are designed to support mathematical development by providing learning opportunities to increase children's informal mathematical knowledge. The intervention consists of a sequence of small-group math activities with concrete manipulatives that teachers implement in the pre-K classroom. The program also includes home math activities in the form of picture strips for parents to use with their children.

The content of the activities is based on developmental research about the nature and extent of early mathematical knowledge (see Geary, 1994 and Ginsburg, Klein, & Starkey, 1998 for early reviews of the research). The curriculum targets a range of pre-K mathematics concepts and skills, including number, operations, geometry, pattern knowledge, and measurement. Units and activities within *Pre-K*

Mathematics are designed to prepare children for each of the clusters of standards included in the Common Core State Standards for Mathematics at kindergarten. They are also explicitly linked to National Council of Teachers of Mathematics (2006) Focal Points. Downward (less challenging) extensions of the mathematics activities are available for children who are not ready for a given activity, and upward (more challenging) extensions are available for children who complete an activity easily.

Teachers attend multi-day professional development workshops in which they learn about the philosophy and key features of the program as well as how to implement the math activities. In addition, *Pre-K Mathematics* employs several implementation tools, and teachers get hands-on experience with them in the workshops. They learn how to keep track of each child's learning over the course of the year, using recording sheets that accompany each math activity and a progress monitoring tool that documents the child's mastery of the math concepts targeted by the curriculum. Teachers also send home weekly math activities (in English or Spanish) for parents to engage with their children. During the workshops, teachers learn how to explain these at-home activities to parents and how to use a parent feedback form to document parents' use of these activities.

2.2. Prior Studies of *Pre-K Mathematics*

The developers conducted early pilot studies of *Pre-K Mathematics* as part of the intervention's development. These studies were either non-experimental or they did not evaluate the intervention as a whole. All yielded positive, statistically significant impact estimates (Starkey & Klein, 2000; Starkey, Klein, & Wakeley, 2004).

In this paper, we focus on four studies of the efficacy and effectiveness of *Pre-K Mathematics*, referring to them here as Study 1 (Klein, Starkey, Clements, Sarama, & Iyer; 2008), Study 2 (Starkey, Klein, & DeFlorio, 2014; Starkey & Klein, 2012), Study 3 (Starkey & Klein, 2014), and Study 4 (Starkey, Klein, & Clarke, 2015). In the rest of this section, we discuss how the current study and the four prior studies differ in terms of the dimensions of scale-up from our framework.

2.2.1. Settings and Study Participants

In terms of settings and study participants, the current study is the largest and most diverse, with unclear implications for effect sizes. The number of pre-K classrooms is 50 percent larger than the number in the next-largest study (140 compared to 94 in Study 2). Although Studies 1 and 2 included pre-K sites in other states as well as in California, the California sites were all located in Northern California (the Bay Area for Study 1, and the greater Sacramento region for Study 2). Study 3 sites were all in the Bay Area and Sacramento regions, whereas Study 4 took place only in the Bay Area and Central Valley of California.

In contrast, the current study took place in all major California regions: the greater Los Angeles area, the Central Valley, the Bay Area, and rural Northern California. The pre-K sites, which were purposively selected, included public pre-K and Head Start programs in urban, suburban, and rural areas with large proportions of low-income families from diverse racial/ethnic backgrounds. The case for absolute scale-up to the state level rests on this heterogeneous sample and the opportunity it provides to replicate results across purposively (instead of randomly) selected sites (Cook, 2014).

[Table 1 about here.]

The sample size of children in this study is nearly twice as large as it is in Study 2 (1,373 vs. 744 at baseline). Studies 1, 3, and 4 were smaller (ranging from 316 to 526 children).[1] Although the majority of Study 2 children were white, the proportions of minority children in the other studies are similar to those in the current study. Being larger, though, the current study has larger numbers of children from a variety of racial/ethnic groups. Hispanic children account for three-fourths (75 percent) of the total study sample. The sample also includes white, African American, Asian, and mixed-race children.

[1] Study 3 examined two interventions: Pre-K Mathematics in the pre-K year, and Pre-Pre-K Mathematics in the year before pre-K. A total of 526 children participated in the study. The study had two treatment conditions: 179 children received Pre-Pre-K Mathematics at age 3 and Pre-K Mathematics at age 4, and 172 received only Pre-K Mathematics at age 4. The remaining 175 made up the control group.

2.2.2. Program Content and Delivery

Our theory predicts that the program content could change as a result of scale-up—for example, intervention developers could improve program content to incorporate lessons learned from prior studies, which would imply an increase in effect sizes, all else equal. The content of the *Pre-K Mathematics* curriculum was largely similar across all four studies. The broad intervention approach and core math activities remained the same: incorporating classroom and home activities, emphasizing teacher training and coaching, and collecting detailed implementation data. In addition, all four studies included mathematical enrichment of the classroom learning environment (math software, teacher-created math centers, or both).

However, across studies, the program developers learned from experience and refined some program details accordingly. For example, the developers learned in this study that some activities turned out to be too long to complete in one week, and they altered the curriculum schedule to allow two weeks for those activities. The developers also dropped some activities and added others to better align the curriculum to the Common Core State Standards for Mathematics, which were being stressed by the national and state Departments of Education. We hypothesize that any influence these incremental improvements would have on effect sizes would be small.

The theory predicts that the quality of program delivery is likely to suffer at larger scales, due to less developer control over the implementation of the intervention. The level of developer control over how *Pre-K Mathematics* was delivered differed across the studies we examine here, with unclear implications for effect sizes. Study 2 used a train-the-trainers model, in which the professional development staff hired by the project (that is, project staff) trained program trainers, who then trained and coached the teachers implementing the intervention. When possible, the developers worked within the preschool programs' existing professional development system for teachers. However, when programs did not have their own professional development staff, they hired outside contractors to train and coach the teachers.

In contrast, Studies 1, 3, and 4 used a direct training model in which project staff trained teachers and coached them. In the current study, the developers used a hybrid model depending on the needs and capacity of the participating preschool programs. In some sites, project staff directly trained and coached teachers, and in others, they used a train-the-trainers approach. These differences in training methods entail less developer control over training and monitoring in the current study than in Studies 1, 3, and 4, but somewhat more control than in Study 2. Although these differences might seem marginal and unlikely to have much influence on effect sizes, if operating alone, they would lead to predicting smaller effects in the present study than in Studies 1, 3, and 4, whereas they would predict larger effects than in Study 2.

2.2.3. Changes in the Counterfactual Condition

Our theory predicts that changes in the counterfactual condition can have negative consequences for effect sizes if the business-as-usual comparison group's achievement has increased over time, which we expect is the case for preschool mathematics. The studies of *Pre-K Mathematics* we examine were conducted over a 12-year span between 2003 and 2015 (publication dates range from 2008 to 2018). All the studies included California preschool sites, though several also included sites in other states (New York and Kentucky). California adopted the Common Core State Standards in 2010, though implementing the curriculum, instructional materials, and assessments based on the standards will take several years (California Department of Education, 2016). Increased awareness of these standards will likely improve the state of mathematical knowledge in comparison groups over time, and thereby reduce effect sizes in the current study compared to past studies.

Within this time span, California also changed its cutoff age for kindergarten entry and instituted transitional kindergarten statewide for eligible 4-year-olds.[2] However, we do not expect this policy change to result in substantial time-varying differences in the state of mathematical knowledge in the

[2] Before the 2012–2013 school year, California state law stipulated that children must turn 5 by December 2 to be able to start kindergarten in that year. A new law (SB 1381, which became effective in the 2012–2013 school year) stipulated a new cutoff date of September 2. To phase in this new age requirement, the state moved the cutoff date back one month each year for three years. The cutoff was November 2 in the 2012–2013 school year, October 2 in 2013–2014, and September 2 in 2014–2015. SB1381 also established transitional kindergarten for students who would turn 5 between September 2 and December 2 (that is, those students who would have been able to start kindergarten under the old law, but could not under the new law) (Mercado-Garcia, Quick, Holod, & Manship, 2013). State transitional kindergarten programs were not included in any of the studies of Pre-K Mathematics.

comparison groups, because transitional kindergarten was offered to only a few 4-year-olds—those who would otherwise have started kindergarten if not for the change in the entry cutoff.

2.2.4. Outcome Measurement Quality

Overall, our theory predicts that less developer control over measurement will decrease reliability and lower effect sizes. The studies of *Pre-K Mathematics* we examine differed in terms of how much the developer controlled the assessment process. In the four prior efficacy and effectiveness studies, research staff hired by the intervention developers trained all the assessors and directly monitored them. In the current study, funding requirements dictated that the developers use an external data collector. An independent research firm carried out the assessment process under contract with the intervention developers. A train-the-trainers model was used for all child outcome measures. That is, project research staff trained the trainers from the external data collection firm, and these trainers then hired and trained the child assessors and oversaw the assessment process.

With regard to the measures themselves, we expect that less well-aligned measures will yield lower effect sizes. The current study as well as the prior studies used math assessments that differed in terms of their alignment with the *Pre-K Mathematics* intervention. All four prior studies had one primary math outcome measure in common: the Child Math Assessment (CMA). The CMA is a researcher-developed measure that is aligned with the math content of the *Pre-K Mathematics* intervention, but it does not use any of the same materials as the intervention. The CMA assesses informal mathematical knowledge across a broad range of skills and concepts including number, arithmetic, space and geometry, measurement, and patterns. Though the CMA is not a norm-referenced test, it has strong psychometric properties (test-retest reliability over a two-week interval is 0.91, Cronbach's alpha over all tasks is 0.90, and convergent validity with the TEMA-3 is 0.74 [Klein et al., 2008]).

Studies 2, 3, and 4 also used the Test of Early Mathematics Ability, Third Edition (TEMA-3; Ginsburg & Baroody, 2003) as a secondary math outcome measure. The TEMA-3 is a norm-referenced test (but not based on a national sample) that measures informal and formal knowledge in number and operations only. It is therefore not as well aligned with the intervention, whose math content in not limited to number and operations, but also includes space and geometry, measurement, and pattern knowledge. In addition, the TEMA-3 includes a large number of items that test formal (that is, symbolic) number knowledge, making this measure less developmentally appropriate for preschool children.

The current study used the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Mathematics Assessment (U.S. Department of Education, 1998–1999) as the primary math outcome measure. The ECLS-B is an item response theory-based, adaptive test that measures a broad range of early math content including number and arithmetic, geometry, measurement, and pattern knowledge. Moreover, it is a norm-referenced test that is based on a purposive national sample. In addition to ECLS-B, the current study used theTEMA-3 as a secondary math measure to capture growth in children's formal number knowledge in the kindergarten year. Overall, the ECLS-B is better aligned with the content of *Pre-K Mathematics* than the TEMA-3, but not as well-aligned as the CMA.

All three math assessments—the CMA, the TEMA-3, and the ECLS-B—have strong psychometric properties, but they differ in terms of their alignment with the *Pre-K Mathematics* intervention. We hypothesize that effect sizes will decrease as alignment decreases, so the CMA should yield the largest effect sizes, followed by the ECLS-B, then the TEMA-3.

2.2.5. Quality of the Evaluation Design and Its Execution

Though interventions selected for scale-up likely have been evaluated in high quality studies, including RCTs, we postulate that the execution of the planned study design will tend to be lower as the size and heterogeneity of a scale-up study increases, yielding lower effect sizes. Because all of the studies described in Table 1 were cluster RCTs, the quality of their planned evaluation study design was similar. Moreover, the execution of the evaluation was also similar: there are no indications that pre-test balance or rates of treatment contamination varied much from one study to another. It is unlikely, therefore, that evaluation design or execution can be responsible for any effect differences by study.

2.2.6. Overall Predictions for Effect Sizes from Prior Efficacy and Effectiveness Studies and the Current Studies

The current study differs from Studies 1–4 in three meaningful ways. It is statewide; it was conducted later, in a context of higher awareness of the importance of mathematics instruction in pre-K;

and there was less developer control over the measurement process. The first difference has unclear implications for effect sizes; the last two, however, suggest that the current study will have smaller effect sizes than the other three.

3. METHODS IN THE SCALE-UP STUDY OF *PRE-K MATHEMATICS*

In this section, we describe the methods we used to estimate the impacts of *Pre-K Mathematics* in the current study, and those we used to compare these estimates to estimates from past studies of the intervention.

3.1. Experimental Design for the Statewide Scale-Up Study

This study used a cluster RCT design in which pre-K classrooms were randomly assigned to the treatment or control condition. *Pre-K Mathematics* was the intervention implemented in the treatment condition, and the control condition entailed business-as-usual instruction in pre-K. Four-year-old children who were eligible to attend kindergarten the following year, spoke either English or Spanish, did not have an identified developmental disability, and had their parents' or guardians' consent to participate were eligible for the study. Consent was obtained before random assignment. Up to 12 children per pre-K classroom were selected (if more than 12 were eligible, 12 were randomly selected to participate). The intervention was implemented across two cohorts. The first began pre-K in the 2013–2014 school year, and the second in the following year.

We used a heterogeneous but purposive sample of pre-K school sites and classrooms for this study. We recruited pre-K school sites from the greater Los Angeles area, the Central Valley, the Bay Area, and rural Northern California. The purposive selection included public pre-K and Head Start programs located in urban, suburban, and rural areas with large proportions of low-income families from diverse racial/ethnic backgrounds who plausibly represent all low-income families in California with a child of pre-K age.

Though random sampling followed by random assignment is the ideal, we did not employ random sampling in the present study; indeed, for practical reasons, it is rare today and not likely to become standard in the future (Cook, 2014). Nor could we weight the sample to approximate the state-level population, because we did not have access to a data-rich profile of all California children attending pre-K. Finally, we did not randomize within racial and ethnic subgroups, so statistical power to detect meaningful treatment-control differences within these subgroups is limited.

3.2. Measures for the Statewide Scale-Up Study

We used the ECLS-B and the TEMA-3 to collect data on math achievement at post-test (spring pre-K). As a pre-test measure, we used the ECLS-B, administered at the beginning of the pre-K year (fall pre-K).[3] We collected other baseline measures to serve as control variables. One was the Test of Preschool Early Literacy (TOPEL), which is individually administered and assesses the early literacy of children ages 3 to 5 (Lonigan, Wagner, Torgesen, & Rashotte, 2007). Others were demographic covariates: age, gender, race/ethnicity, and language.

We collected two implementation measures during the school year. To assess fidelity, local trainers made visits to teachers in the treatment classrooms. They observed as a teacher conducted a small-group math activity and gave the teachers feedback afterward about any departures from fidelity. The second measure was the Early Mathematics Classroom Observation (EMCO), an observation tool used to determine the nature and amount of mathematics instruction that preschool teachers provided in their classrooms. For each teacher-participant activity involving mathematical content, trained observers recorded the type of mathematical content, number of children present, and the duration of the activity. This provided data on the number of minutes of math instruction, on average, to a child during an observation session.

3.3. Sample for the Statewide Scale-Up Study

At the time of random assignment, the sample consisted of 1,373 children (687 in the treatment group; 686 in the control group). They came from 140 pre-K classrooms (70 treatment; 70 control) within 106 pre-K school sites and 10 school districts. There were 17 pre-K classrooms in the Bay Area, 31 in

[3] We did not assess children on the TEMA-3 at pre-test, because its principal purpose was to supplement our assessment of children's formal number knowledge in kindergarten. .

rural Northern California, 13 in the rural Central Valley, and 79 in various parts of Southern California, including the greater Los Angeles area. Overall, 18 percent of pre-K classrooms were in urban areas, 51 percent in suburban areas, and 31 percent in towns or rural areas.

These pre-K classrooms comprise an ethnically and linguistically diverse population of low-income families. Seventy-five percent were Hispanic, 13 percent were white, 6 percent were African American, 4 percent were of mixed race, and 2 percent were Asian. Most of the sample were exclusively English speakers (68 percent); for 25 percent, Spanish was the dominant language; and 7 percent spoke both English and Spanish. On average, children were 4.4 years old at baseline (fall of the pre-K year), and 48 percent were male (Table 2).

[Table 2 about here]

The analytic sample for the ECLS-B consisted of 1,313 children (653 in the treatment group and 660 in the control) within 70 treatment and 70 control classrooms. For the TEMA-3, which was not administered in five classrooms at post-test, the sample was somewhat smaller: 1,256 children (621 treatment, 635 control) within 135 classrooms (67 treatment, 68 control). Overall attrition was low: 4.4 percent for the ECLS-B, and 8.5 percent for the TEMA-3. Differential attrition was also low: 1.2 percent for the ECLS-B, and 2.2 percent for the TEMA-3.

Table 3 shows that the final treatment and control samples used in the analysis were well balanced on the available variables, showing no significant differences on pre-test or baseline demographic characteristics. Even though there were treatment and control classrooms within the same school, classroom observations, during which the EMCO was administered, revealed no borrowing of the math curriculum by the control classrooms. When combined with all the information presented above, it looks as though the experimental design was adequately implemented.

[Table 3 about here.]

3.4. Hypotheses and Impact Models for the Statewide Scale-Up Study

The null hypothesis for testing whether the pre-K math program is effective at the state level is that the treatment and control groups do not differ on their post-test math performance. The counter-hypothesis is that the group exposed to *Pre-K Mathematics* does better.

Because scale-up entails greater heterogeneity in the subpopulations of persons and settings sampled, we also test the null hypotheses that effect sizes do not depend on (1) race/ethnicity; (2) prior knowledge of math—that is, whether children had higher or lower performance at pre-test; and (3) urbanicity; that is, residence in urban, suburban, or rural sites within California. The counter-hypotheses are that intervention effects are heterogeneous across children and settings.

We present regression-adjusted means for the treatment and control groups from a hierarchical linear model in which children are nested within pre-K classrooms, the unit of random assignment. The model includes the full set of child-level covariates (ECLS-B and TOPEL pre-tests, age, gender, race/ethnicity, language, and cohort). We then calculate effect sizes using Hedges' g formula:

$$(2)\quad g = \frac{\widehat{mean}_T - \widehat{me\quad}\ _C}{SD_{pooled}}$$

where $\widehat{mean}_T$ equals the adjusted treatment group mean, $\widehat{mean}_C$ equals the adjusted control group mean, and SD_{pooled} is the pooled SD. We use the following formula to calculate the pooled SD:

$$(3)\quad SD_{pooled} = \sqrt{\frac{(N_T - 1)SD_T^2 + (N_C - 1)SD_C^2}{N_T + N_C - 2}}$$

as the difference between the regression-adjusted treatment and control group post-test means, divided by the unadjusted, pooled treatment and control group standard deviation.

3.5. Methods for Comparing Effect Sizes across Studies

As a simple way to compare effect sizes across studies, we examine how the authors' preferred effect size calculations differ across studies by plotting the obtained effect sizes from Studies 1–4 and the current study against the year in which pre-K took place for each study. We plot these effect sizes for all math outcomes used in these studies: as discussed, Study 1 used the CMA; Studies 2, 3 and 4 used both the CMA and TEMA-3; and the current study used the TEMA-3 and the ECLS-B.

However, these different measures pose a problem for comparing effect size differences across studies that have grown larger, more heterogeneous, and less well controlled. The tests vary in many ways that might influence effect sizes, but especially in how well they are aligned with program content. To remove one source of variation when comparing effect sizes—different outcome measures—we will also compare effect sizes across the four studies that used the TEMA-3 (Studies 2, 3, and 4 and the current evaluation). This reduces the length of the time span examined and relies on the least well-aligned outcome. Even so, the time period and math measure should be enough to see if effect sizes trend downward due to reasons other than different outcome measures.

Past studies also differ in the causal quantity estimated, with some computing intent-to-treat (ITT) estimates and others calculating treatment-on-the-treated (TOT) estimates. To hold the estimand constant, we calculated unadjusted ITT effect sizes for each study that used the TEMA-3, using for this purpose the unadjusted treatment and control group post-test means and standard deviations for the ITT sample:

$$(4)\quad ITT\ effect\ size = \frac{Unadjusted\ treatment\ group\ mean - Unadjus\qquad comparison\ group\ mean}{Unadjusted\ control\ group\ standard\ deviation}$$

Such details were generously provided to us by study authors when they were not already available from research reports.

4. RESULTS FOR THE STATEWIDE SCALE-UP STUDY OF *PRE-K MATHEMATICS*

4.1. Average Effects of the *Pre-K Mathematics* Intervention

Pre-K Mathematics had positive and significant effects on the math achievement of pre-kindergartners as measured by the ECLS-B and the TEMA-3 at the end of the pre-K year. Table 4 presents the regression-adjusted post-test means and unadjusted, pooled post-test standard deviations from the model described earlier, for the total scale-up effect. The program was clearly effective, as prior studies also showed: the effect size was 0.30 for the ECLS-B and 0.23 for the TEMA-3, with a *p*-value of less than 0.01 for both outcome measures.

[Table 4 about here]

4.2. Interactions of the Intervention with Child and Site Characteristics

In post-hoc subgroup analyses, we found that treatment effect estimates were positive in sign for all racial/ethnic groups, and they did not reliably differ across the groups. The group-specific effect sizes on the ECLS-B ranged from 0.24 for white and Asian children to 0.59 for African American children; on the TEMA-3, they ranged from 0.12 for white children to 0.55 for African American children (Table 5). Sample sizes varied considerably by group, and there were no reliable interactions of treatment and racial/ethnic groups within the power limits imposed by the sample sizes.[4] All indications are that each population group benefitted from the intervention, and the consistently positive causal signs suggest that none was negatively affected.

[Table 5 about here]

We also examined average treatment effects by pre-test performance and by urbanicity. There is no consistent evidence that treated children who scored lower on the pre-test did appreciably better or worse over time than their higher-scoring counterparts (Table 6). In addition, effect sizes do not vary systematically among pre-K sites located in cities, suburbs, or towns/rural areas (Table 7). Statistical analyses confirm that neither the initial math score nor the three location categories interact with treatment to affect average differences in effect sizes.[5]

[Tables 6 and 7 about here]

[4] To test this, we regressed the ECLS-B scale score and the TEMA-3 at wave 2 on interactions between each racial/ethnic category and treatment (four interactions total) and included the other baseline control variables (ECLS-B and TOPEL pretests, age, gender, language, and cohort). We tested whether the coefficients on the interaction terms were equal and were unable to reject the null hypothesis that they were equal (p = 0.559 for the ECLS-B and p = 0.433 for the TEMA-3).

[5] We tested this in the same way we tested differences in treatment effects across racial/ethnic categories. In all cases, we were unable to reject the null hypothesis at conventional significance levels that the coefficients on the relevant interaction terms were equal. For the test of the equality of treatment-ECLS-B pretest quintile interaction coefficients, p = 0.267 for the ECLS-B and p = 0.097 for the TEMA-3. For the test of equality of treatment-urbanicity interaction coefficients, p = 0.635 for the ECLS-B and p = 0.961 for the TEMA-3.

5. COMPARING EFFECT SIZES ACROSS STUDIES

Figure 2 plots study-specific effect sizes against the year in which spring pre-K took place for Studies 1–4 and for the current evaluation of *Pre-K Mathematics*. A downward trend is clearly apparent: effect sizes decrease over time.

[Figure 2 about here.]

Of course, the study date is correlated with many other things besides time, such as the math measure that was used. The figure also shows the different math outcome measures used in these studies: the CMA, the TEMA-3, and the ECLS-B. This reveals that effect sizes are larger with better-aligned measures. In each pre-K year in which children were assessed using both the CMA and the TEMA-3—that is, 2008, 2011, and 2014—CMA effect sizes were larger than TEMA-3 effect sizes. In the current study, effect sizes were also larger with the better-aligned measure: the ECLS-B is better aligned than the TEMA-3 (though not as well aligned as the CMA is), and the ECLS-B yielded a larger effect size than the TEMA-3.

Table 8 shows what happens when effect sizes are calculated in exactly the same way across studies, but limited to the more recent studies using the TEMA-3—Studies 2, 3, and 4. A temporal pattern of reduced average effect sizes continues to be apparent, but it is now more modest, and ranges from a high of 0.45 standard deviations in Study 3 to a low of 0.20 in the present study.[6]

[Table 8 about here.]

The computation of effect sizes includes consideration of the differences in the sample sizes of each study, thanks to the standard deviation included in the denominator of Formula 1. Nonetheless, we might compute a rough measure of "total" program impact to include both program performance and demonstrated program reach by multiplying the average effect size by the number of children in a study.

[6] To compare effect sizes across studies that used the TEMA-3, the effect sizes in Table 8 are calculated differently from the effect sizes in Table 4, the main results table. Specifically, Table 4 reports effect sizes calculated using adjusted treatment-control mean differences, and Table 8 reports effect sizes calculated using unadjusted mean differences.

By this metric, the study with the largest effect size, Study 2, has a relatively small total program impact due to its modest sample size, and the study with the smallest effect size, the current evaluation, has the largest total program impact of the four studies in Table 8.

6. DISCUSSION

This paper presented a case study of absolute scale-up of a pre-K math program to the state level in California. It was not possible to enumerate all pre-K students in the state, or even all pre-K students from low-income families. So it was not possible either to select students at random before randomly assigning them to the *Pre-K Mathematics* intervention or to weight the sample we achieved to better approximate the relevant population on attributes measured in both the population and the sample. Instead, our rationale for state-level generalization is predicated on the heterogeneity of the achieved sample in terms of the location of children within the state and of their demonstrated variation in many other attributes that past research has shown to be related to math gains. Such a sampling plan is feasible and widely practiced across all the social and natural sciences.

With this proviso in mind, we tested whether the positive evaluation results that were demonstrated in earlier, smaller, and more homogeneous studies continued to be observed when essentially the same intervention was implemented on a larger and more heterogeneous scale. We showed that *Pre-K Mathematics* had consistently positive effects, and that these effects did not demonstrably differ by the racial/ethnic background or pre-test performance of children, or by the urbanicity of the settings.

The present study also set out to test whether scale-up lowers achievement gains relative to those found in earlier research. Past evaluations of *Pre-K Mathematics* over more than 10 years show a decreasing trend over time, consistent with the hypothesis that larger, more heterogeneous, and less controlled studies tend to yield smaller results. However, the math outcome measures also varied over this period. When we use the same TEMA-3 outcome to contrast study results, we limit the analysis to the more recent past, and the resulting historical trend is less certain and less steep, but still evident.

Hence, we find support for our conjecture, based on the framework of dimensions of scale-up that we developed in this paper, that scale-up is associated with smaller effect sizes—as studies got larger, more heterogeneous, and less controlled, and used less well-aligned outcome measures, effect sizes tended to decrease. Of course, it is possible to imagine scenarios in which budgets increase dramatically with study size and experience has accumulated to counteract the forces that diminish effect sizes. However, we judge it likely that larger scales will often entail lower costs per study unit sampled, because reducing per-unit costs is one rationale for conducting scale-up studies.

In other words, scale matters. But how it matters cannot be interpreted solely in terms of lower average effect sizes over time. Scale-up also tends to increase the size and heterogeneity of the group for which lower effects are shown. Construing total program impact to include both program performance and demonstrated program reach would mean that interpreting the policy implications of modestly smaller effect sizes would need to be conditional on the higher numbers and greater heterogeneity over which effects were demonstrated. In other words, the study's higher internal validity of smaller findings would be balanced against its demonstrably higher external validity.

Other early childhood education programs have claimed positive impacts on some outcomes assessed in adulthood instead of early childhood (Monahan, Thomas, Paulsell, & Murphy, 2015). If that is the emerging standard for assessing the importance of pre-K results, then children will have to build on the gains observed here. But the mechanisms for capitalizing on pre-K gains are not known yet, and perhaps the greatest challenge to pre-K education is to discover which mechanisms translate early gains in literacy and numeracy into more years of schooling, better employment, less involvement in the criminal justice system, and more stable adult relationships. A modest but broadly distributed gain in math achievement within a state or nation is more meaningful if it can be linked to changes in some broad set of cognitive and non-cognitive skill changes that might plausibly maintain or transform early achievement gains.

The present study also has implications for how we think about the external validity of causal relationships. One formulation of causal generalization (Cook, 2014), based on work by Cronbach et al.

(1980) and Mackie (1975), frames external validity in terms of two dimensions. The first involves five domains in which generalization is explicitly or implicitly sought: the target populations of (1) times, (2) persons, (3) settings, (4) treatments, and (5) outcomes.

The second dimension of external validity involves different functions of causal generalization. One function concerns the populations and constructs that the study-level samples and measures stand for. The key question here is: What do these sampling particulars "represent"?

The second causal generalization function uses these same sampling particulars to generalize a causal connection to cause-and-effect constructs (that is, treatments and outcomes) and to populations of persons, settings, and times that manifestly differ from what the sampling particulars are judged to represent. The key question here is: How can we extrapolate from the obtained sampling specifics to constructs and populations that are manifestly different from what the sampling specifics represent? This latter causal generation function equates external validity with "extrapolation" instead of "representation," and it probably better describes what confronts potential users of causal information. They need to know: How can the information from causal studies be applied to the program for which I am responsible, given that any changes I make (1) will be in the future and not the past, (2) will apply to the people I work with and not those studied in the past, (3) will be implemented in the setting where I work and not in the sample of settings examined earlier, (4) will have to apply to the program updates and adaptations I have to make or want to make, and (5) will also have to apply to the way I want to measure the effect, as opposed to the ways it was done before?

The present study is part of a sequence of experiments evaluating ostensibly the same pre-K mathematics program. All the studies included children of similar ages whose families generally had low incomes. But the attributes of the children differed in other ways, including observed race/ethnicity, pre-test scores, and many other unobserved attributes that developmental psychologists would find important as possible moderators of treatment effects. Yet, despite this heterogeneity in student attributes, positive effects were routinely obtained over time.

The settings are also different across the studies, including the states and urbanicity of the geographic regions in which the evaluations took place, the kinds of school districts, and type of the pre-K program (state preschool and Head Start). The time period is also different, lasting over 10 years from the first to the last study. Yet positive effects were routinely found despite these secular changes. Also unique over time are the program variants we sampled, for the program theory marginally changed over time, as did the implementation of the program particulars and the visibility of the developer. Finally, several different math outcome measures have been used, each showing positive effects. There is no doubt, then, that the sampling particulars in this synthesis of studies indicate that *Pre-K Mathematics* is robustly effective across many sources of variation in persons, settings, time, and ways of operationalizing the cause and effect.

But robust effectiveness across multiple sources of sampled heterogeneity is not the same as formal generalization to specifically named populations or cause and effect constructs. Random sampling is best for this, but it is rarely used in causal research and does not lend itself to generalizing to times or to cause-and-effect constructs. To improve representativeness, we might have weighted sample details to better represent known population details, but this is impossible because no enumeration of the statewide pre-K population is available. Instead, we used a strategy based on demonstrating effectiveness across multiple sources of sampled heterogeneity, this being the current strategy for generalizing causal connections in the social and natural sciences. The claim we make is that the sources of heterogeneity we sampled *across studies* is unusually rich in attributes that are plausibly related to math gains, and that such gains were routinely observed.

Extrapolating a causal relationship to unobserved populations and constructs is even more complicated. When researchers directly examine extrapolation, it is usually in the context of persons rather than settings, times, causes, or effects. The researchers in question then model how a causal relationship would change if the personal attributes sampled were differently weighted to better reflect the intended population. This approach is limited, however, because it cannot deal with novel attributes of the target population that are not measured in the obtained sample.

We believe that extrapolation is best promoted at present by the strategy adopted here. Within studies, it emphasizes the purposive but deliberately heterogeneous sampling of persons and sites and then emphasizes testing the generalization of the causal connection across these sources of heterogeneity. More important is that it also emphasizes purposive sampling between studies, for this will usually create even more heterogeneity than a single study achieves. The consistency of effect sizes can then be ascertained over this even greater heterogeneity in persons, settings, times, and cause-and-effect variants, defining consistency in terms either of causal signs (achieved here), or of statistically significant differences (also achieved here), or of effect sizes (not demonstrated here because effect sizes tended to decrease with time and increases in study size, as predicted by our framework).

This operative rationale for extrapolation is admittedly inductive. It assumes that a robustly replicated causal relationship across the multiple sources of variation examined to date raises the odds that the same causal relationship will continue to be inferred in future populations and with future modifications to the cause-and-effect operations. Of course, there can be no guarantee of this. Because the future is never identical to the past, this rationale for extrapolation requires the pragmatic conditional that a heterogeneously demonstrated causal relationship will continue to hold *until proven otherwise*. Ontologically, it seems unlikely that causal relationships are universal; likely, all of them will be conditional on something. Realistically, we cannot wait until all the relevant causal contingencies have been identified. We have to choose to act, and our proposal is to wait to act until a causal relationship has been heterogeneously tested and shown to be robust, and then to act as though it were universally true while remaining mindful that later experiences may help identify the contingencies under which the given causal relationship is indeed true.

To seek to generalize causal relationships only to persons represents an advance, but it does not directly speak to the other four domains that are intrinsically embedded in any causal claim. It also frames external validity more in the service of how a sample of persons represents the researcher's intended population than in how the sample might facilitate extrapolation to targets with attributes that are not included among the sample details. Extrapolating a causal relationship is admittedly more complex than representing what the sampling details in a causal study represent. Moreover, it is less amenable to conceptualization within the conventional sampling theory perspective that dominates recent thinking about external validity. Our concern is that approaching external validity and causal generalization from this perspective will lead to our seeing only part of the elephant. We suggest that the purposive but heterogeneous sampling that multiple tests of the same evolving program make possible is better for causal extrapolation, and probably better still when the more recent studies have larger and more heterogeneous sampling frames and less researcher and developer control over study details—as in the scale-up context examined here.

REFERENCES

California Department of Education (2016). Common Core State Standards (CCSS) answers to frequently asked questions (FAQs). Retrieved from http://www.cde.ca.gov/re/cc/ccssfaqs.asp.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., ... & Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Cook, T. D. (2014). Generalizing causal knowledge in the policy sciences: External validity as a task of both multi-attribute representation and multi-attribute extrapolation. *Journal of Policy Analysis and Management, 33*(2), 527–536.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, …, & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*(3), 151–175.

Geary, D. C. (1994). *Children's mathematical development: Research and practical applications*. Washington, DC: American Psychological Association.

Ginsburg, H. P. & Baroody, A. J. (2003). *Test of Early Mathematics Ability–Third edition*. Austin, TX: Pro-Ed.

Ginsburg, H. P., Klein, A., & Starkey, P. (1998). The development of children's mathematical thinking: Connecting research with practice. In W. Damon, I. E. Sigel, & K. A. Renninger (Eds), *Handbook of child psychology* (pp. 401–476). Hoboken, NJ: John Wiley & Sons.

Glennan, T. K., Bodilly, S. J., Galegher, J., & Kerr, K. A. (2004). Expanding the reach of education reforms. *Perspectives from leaders in the scale-up of educational interventions.* Santa Monica, CA: RAND.

Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*(7), 893–926.

Earle, J., Maynard, R., Neild, R. C., Easton, J. Q., Ferrini-Mundy, J., Albro, E., & Winter, S. (2013). Common guidelines for education research and development. Washington, DC: U.S. Department of Education Institute of Education Sciences and National Science Foundation.

Klein, A. & Starkey, P. (2002). *Pre-K Mathematics curriculum*. Glenview, IL: Scott Foresman.

Klein, A. & Starkey, P. (2004). *Scott Foresman–Addison Wesley Mathematics: Pre-K*. Glenview, IL: Pearson Scott Foresman.

Klein, A., Starkey, P., Clements, D., Sarama, J. & Iyer, R. (2008). Effects of a pre-kindergarten mathematics intervention: A randomized experiment. *Journal of Research on Educational Effectiveness, 1,* 155-178.

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research*, *45*(5), 626–629.

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

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *TOPEL: Test of preschool early literacy*. Austin, TX: Pro-Ed.

McDonald, S. K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher*, *35*(3), 15–24.

Mackie, R. R. (1975). Toward a criterion of training device acceptance. In *Proceedings of the Human Factors Society Annual Meeting, 19*(1). Los Angeles, CA: SAGE Publications.

Mercado-Garcia, D., Quick, H. E., Holod, A., & Manship, C. (2013). Transitional kindergarten in California: Initial findings from the first year of implementation. Washington, DC: American Institutes for Research.

Monahan, S., Thomas, J., Paulsell, D., and Murphy, L. (2015). "Learning about infant and toddler early education services (LITES): a systematic review of the evidence." Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.

National Council of Teachers of Mathematics. (2006). Curriculum focal points pre-k. Reston, VA: National Council of Teachers of Mathematics.

Preschool Curriculum Evaluation Research Consortium (2008). *Effects of preschool curriculum programs on school readiness* (NCER 2008–2009). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Starkey, P., & Klein, A. (2000). Fostering parental support for children's mathematical development: An intervention with Head Start families. *Early Education and Development, 11*(5), 659–680.

Starkey, P., & Klein, A. (2012). Scaling up the implementation of a pre-kindergarten mathematics curriculum in public preschool programs (IES Grant R305K050004): Final report.

Starkey, P. & Klein, A. (2014). Closing the SES-related gap in young children's mathematical knowledge (IES Grant R305A080188): Final report.

Starkey, P, Klein, A., & Clarke, B. (2015). A Randomized Study of the Efficacy of a Two-Year Mathematics Intervention for At-Risk Pre-Kindergarten and Kindergarten Students. Paper presented at the 2015 Institute of Education Sciences Principal Investigators Meeting.

Starkey, P., Klein, A., & DeFlorio, L., (2014). Promoting math readiness through a sustainable mathematics intervention. In Bierman, K.L. & Boivin, M. (Eds.) *Promoting school readiness and early learning: the implications of developmental research for practice*. New York: Guilford.

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly. 19,* 99–120.

Tipton, E., Hallberg, K., Hedges, L.V., and Chan, W. (2016). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review, 41*(5), 472–505.

U.S. Department of Education, National Center for Education Statistics. (1998–1999). *ECLS-B mathematics assessment.* Washington, DC: Author.

What Works Clearinghouse. (2013). *Pre-K mathematics with DLM early childhood express math*. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from http://whatworks.ed.gov.

Yamey, G. (2011). Scaling up global health interventions: a proposed framework for success. *PLoS Med*, *8*(6), e1001049.

TABLES AND FIGURES

Table 1. Current and past studies of *Pre-K Mathematics*

Study	Settings and study participants	Program content and delivery	Changes in the counterfactual condition	Measurement quality	Evaluation design
Study 1 (Klein, Starkey, Clements, Sarama, & Iyer, 2008)	• 40 classrooms in CA and NY • Urban areas • 316 children at baseline • 53 percent African American, 22 percent white, 22 percent Hispanic	• Program content: Classroom and home components of *Pre-K Mathematics*; DLM Express math software for classroom use • Program delivery: Direct teacher training	• Spring pre-K took place in 2003	• CMA • Direct assessor training and monitoring	• Cluster RCT
Study 2 (Starkey, Klein, & DeFlorio, 2014; Starkey & Klein, 2012)	• 94 classrooms in CA and KY • Urban and rural areas • 744 children at baseline • All low-income • 52 percent white, 18 percent Hispanic, 17 percent African American	• Program content: Classroom and home components of *Pre-K Mathematics* • Program delivery: Train-the-trainer model	• Spring pre-K took place in 2007 and 2008 • Business-as-usual teaching practices in control group	• CMA, TEMA-3 • Direct assessor training and monitoring	• Cluster RCT
Study 3 (Starkey & Klein, 2014)	• 63 classrooms in northern CA (Bay Area and Sacramento) • Urban, suburban, and rural areas • 526 children at baseline (347 in relevant intervention conditions) • 58 percent Hispanic, 18 percent African American, 14 percent mixed race or other	• Program content: Classroom and home components of *Pre-K Mathematics* • Program delivery: Direct teacher training	• Spring pre-K took place in 2011 • Business-as-usual teaching practices in control group	• CMA, TEMA-3 • Direct assessor training and monitoring	• Cluster RCT
Study 4 (Starkey, Klein, & Clarke, 2015)	• 41 classrooms in CA (Bay Area and Central Valley) • Urban, suburban, and rural areas • 389 children at baseline • All low-income • 76 percent Hispanic, 8 percent mixed race or other, 7 percent white	• Program content: Classroom and home components of *Pre-K Mathematics* • Program delivery: Direct teacher training	• Spring pre-K took place in 2014 • Business-as-usual teaching practices in control group	• CMA, TEMA-3 • Direct assessor training and monitoring	• Cluster RCT

Study	Settings and study participants	Program content and delivery	Changes in the counterfactual condition	Measurement quality	Evaluation design
The current study	• 140 classrooms throughout CA • Urban, suburban, and rural areas • 1373 children at baseline • All low-income • 76 percent Hispanic, 11 percent white, 6 percent African American	• Program content: Classroom and home components of *Pre-K Mathematics*; content explicitly tied to K Common Core standards. • Program delivery: Hybrid of a train-the-trainer model and direct teacher training	• Spring pre-K took place in 2014 and 2015 • Business-as-usual teaching practices in control group	• ECLS-B, TEMA-3 • External data collector	• Cluster RCT

CMA = Child Math Assessment; ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; K = kindergarten; RCT = randomized controlled trial; TEMA-3 = Test of Early Mathematics Ability, Third Edition.

Table 2. Descriptive statistics of baseline, ECLS-B, and TEMA-3 analytic samples

	Baseline sample (wave 1) (N = 1373)		ECLS-B analytic sample (wave 2) (N = 1313)		TEMA-3 analytic sample (wave 2) (N = 1256)	
Baseline characteristic	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**
ECLS-B scale score	19.90	0.28	19.87	0.29	19.82	0.30
TOPEL	10.73	0.34	10.71	0.35	10.65	0.36
Age	4.44	0.01	4.45	0.01	4.44	0.01
Male	0.48	0.01	0.48	0.01	0.48	0.01
Hispanic	0.75	0.02	0.75	0.02	0.77	0.02
White	0.13	0.02	0.13	0.02	0.11	0.02
African American	0.06	0.01	0.05	0.01	0.06	0.01
Mixed/other race	0.04	0.01	0.04	0.01	0.04	0.01
Asian	0.02	0.01	0.02	0.01	0.02	0.01
English	0.68	0.02	0.67	0.02	0.66	0.02
Spanish	0.25	0.02	0.26	0.02	0.27	0.02
English and Spanish	0.07	0.01	0.07	0.01	0.07	0.01

Source: Direct child assessments.

Note: Results adjusted for clustering using a two-level model: children within pre-K classrooms.

ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; SE = standard error; TEMA-3 = Test of Early Mathematics Ability, Third Edition; TOPEL = Test of Preschool Early Literacy.

Table 3. Baseline equivalence on key child-level characteristics, ECLS-B and TEMA-3 analytic samples

	ECLS-B analytic sample						TEMA-3 analytic sample					
	Treatment (N = 653)		Control (N = 660)				Treatment (N = 621)		Control (N = 635)			
Baseline characteristic	Mean	SE	Mean	SE	MD	*p*-value	Mean	SE	Mean	SE	MD	*p*-value
ECLS-B scale score	19.81	0.41	19.92	0.41	-0.12	0.844	19.70	0.43	19.94	0.43	-0.24	0.692
TOPEL	10.55	0.49	10.86	0.49	-0.31	0.656	10.45	0.51	10.85	0.51	-0.39	0.584
Age	4.45	0.01	4.44	0.01	0.01	0.500	4.45	0.01	4.44	0.01	0.01	0.621
Male	0.49	0.02	0.48	0.02	0.01	0.852	0.48	0.02	0.48	0.02	0.00	0.899
Hispanic	0.76	0.03	0.75	0.03	0.02	0.714	0.79	0.03	0.76	0.03	0.03	0.563
White	0.13	0.03	0.12	0.03	0.01	0.806	0.11	0.03	0.11	0.03	0.00	0.987
African American	0.05	0.01	0.05	0.01	0.00	0.877	0.06	0.01	0.05	0.01	0.00	0.829
Mixed/other race	0.04	0.01	0.04	0.01	-0.01	0.597	0.03	0.01	0.04	0.01	-0.01	0.303
Asian	0.01	0.01	0.03	0.01	-0.02	0.104	0.01	0.01	0.04	0.01	-0.02	0.111
English	0.67	0.03	0.68	0.03	-0.01	0.798	0.65	0.03	0.67	0.03	-0.02	0.626
Spanish	0.25	0.03	0.26	0.03	0.00	0.937	0.27	0.03	0.26	0.03	0.01	0.897
English and Spanish	0.08	0.02	0.06	0.02	0.01	0.560	0.08	0.02	0.06	0.02	0.02	0.530

Source: Direct child assessments.

Note: Results adjusted for clustering using a two-level model: children within pre-K classrooms.

ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; MD = mean difference; SE = standard error; TEMA-3 = Test of Early Mathematics Ability, Third Edition; TOPEL = Test of Preschool Early Literacy.

Table 4. Main results, ECLS-B and TEMA-3 post-tests

	Treatment			Control				
Mathematics assessment	N	Adjusted mean	Unadjusted SD	N	Adjusted mean	Unadjusted SD	ES	*p*-value
ECLS-B	653	30.84	5.48	660	29.09	6.24	0.30***	0.000
TEMA-3	621	16.07	7.39	635	14.33	7.51	0.23***	0.000

Source: Direct child assessments.

Note: Results adjusted for clustering using a two-level model: children within pre-K classrooms. We used the following baseline control variables: ECLS-B and TOPEL pre-tests, age, gender, race/ethnicity, language, and cohort. We calculated effect sizes using Hedges' *g* formula:

$$\frac{\widehat{mean}_T - \widehat{mean}_C}{SD_{pooled}}$$

where $\widehat{mean}_T$ equals the adjusted treatment group mean, $\widehat{mean}_C$ equals the adjusted control group mean, and SD_{pooled} is the pooled SD. We used the following formula to calculate the pooled SD:

$$SD_{pooled} = \sqrt{\frac{(N_T - 1)SD_T^2 + (N_C - 1)SD_C^2}{N_T + N_C - 2}}$$

where N_T equals the treatment group sample size, N_C equals the control group sample size, SD_T equals the unadjusted treatment group SD, and SD_C equals the unadjusted control group SD.

We also estimated the impact on the ECLS-B for a treatment-on-the-treated (TOT) sample. The TOT sample consists of treatment children who received at least 75 percent of dosage in the pre-K year, and control children who attended pre-K at least 75 percent of the time ($N_T = 575$; $N_C = 588$). The TOT ES is 0.31 ($p < 0.01$).

***Significantly different from zero at the 0.01 level, two-tailed test.

ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; ES = effect size; SD = standard deviation; TEMA-3 = Test of Early Mathematics Ability, Third Edition.

Table 5. Results by racial/ethnic group, ECLS-B and TEMA-3 post-tests

	Treatment			Control						
Racial/ ethnic group	N	Ad-justed mean	Un-adjusted SD	N	Ad-justed mean	Un-adjusted SD	Impact estimate	SE of impact estimate	ES	*p*-value
ECLS-B post-test										
Hispanic	500	30.26	5.52	501	28.56	6.13	1.70	0.30	0.29***	0.000
White	85	32.64	5.04	75	31.32	5.94	1.32	0.62	0.24**	0.034
African American	36	32.54	5.14	33	29.13	6.36	3.41	0.97	0.59***	0.000
Mixed/other race	24	32.54	4.84	29	30.90	6.76	1.64	0.93	0.27*	0.076
Asian	8	33.24	4.01	22	31.35	6.82	1.89	1.55	0.30	0.223
TEMA-3 post-test										
Hispanic	490	15.32	7.27	493	13.68	7.23	1.63	0.40	0.22***	0.000
White	69	18.52	7.56	61	17.62	6.89	0.90	0.79	0.12	0.253
African American	36	19.16	7.78	33	14.69	8.51	4.47	1.21	0.55***	0.000
Mixed/other race	18	17.89	6.80	26	15.46	7.53	2.43	1.41	0.34*	0.085
Asian	8	20.30	4.90	22	18.98	9.78	1.33	2.41	0.15	0.582

Source: Direct child assessment.

Note: Results adjusted for clustering using a two-level model: children within pre-K classrooms. We used the following baseline control variables: ECLS-B and TOPEL pre-tests, age, gender, language, and cohort. We calculated effect sizes using Hedges' g formula:

$$\frac{\widehat{mean}_T - \widehat{mean}_C}{SD_{pooled}}$$

where $\widehat{mean}_T$ equals the adjusted treatment group mean, $\widehat{mean}_C$ equals the adjusted control group mean, and SD_{pooled} is the pooled SD. We use the following formula to calculate the pooled SD:

$$SD_{pooled} = \sqrt{\frac{(N_T - 1)SD_T^2 + (N_C - 1)SD_C^2}{N_T + N_C - 2}}$$

where N_T equals the treatment group sample size, N_C equals the control group sample size, SD_T equals the unadjusted treatment group SD, and SD_C equals the unadjusted control group SD.

*Significantly different from zero at the 0.1 level, two-tailed test.

**Significantly different from zero at the .05 level, two-tailed test.

***Significantly different from zero at the 0.01 level, two-tailed test.

ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; ES = effect size; SD = standard deviation; TEMA-3 = Test of Early Mathematics Ability, Third Edition.

Table 6. Results by quintile of pre-test performance, ECLS-B and TEMA-3 post-tests

	Treatment			Control						
Quintile of ECLS-B pre-test	N	Ad-justed mean	Unadjusted SD	N	Ad-justed mean	Unadjusted SD	Impact estimate	SE of impact estimate	ES	*p*-value
					ECLS-B post-test					
1	133	24.86	4.79	130	22.54	5.94	2.32	0.70	0.43***	0.001
2	132	28.79	3.95	132	26.51	4.67	2.28	0.55	0.53***	0.000
3	133	31.20	3.83	133	29.13	4.41	2.06	0.49	0.50***	0.000
4	124	33.13	3.12	135	31.82	3.24	1.31	0.38	0.41***	0.001
5	131	36.31	3.47	130	35.46	3.38	0.85	0.40	0.25**	0.035
					TEMA-3 post-test					
1	131	9.28	4.50	126	8.00	4.97	1.28	0.62	0.27**	0.039
2	126	13.12	4.89	127	10.96	4.82	2.16	0.67	0.44***	0.001
3	125	15.18	4.96	125	13.54	4.71	1.65	0.59	0.34***	0.005
4	115	18.86	5.47	133	15.92	5.59	2.94	0.70	0.53***	0.000
5	124	24.31	6.55	124	23.75	6.33	0.56	0.71	0.09	0.433

Source: Direct child assessment.

Note: Results adjusted for clustering using a two-level model: children within pre-K classrooms. We used the following baseline control variables: ECLS-B and TOPEL pre-tests, age, gender, race/ethnicity, language, and cohort. We calculated effect sizes using Hedges' g formula:

$$\frac{\widehat{mean}_T - \widehat{mean}_C}{SD_{pooled}}$$

where $\widehat{mean}_T$ equals the adjusted treatment group mean, $\widehat{mean}_C$ equals the adjusted control group mean, and SD_{pooled} is the pooled SD. We used the following formula to calculate the pooled SD:

$$SD_{pooled} = \sqrt{\frac{(N_T - 1)SD_T^2 + (N_C - 1)SD_C^2}{N_T + N_C - 2}}$$

where N_T equals the treatment group sample size, N_C equals the control group sample size, SD_T equals the unadjusted treatment group SD, and SD_C equals the unadjusted control group SD.

*Significantly different from zero at the 0.1 level, two-tailed test.

**Significantly different from zero at the 0.05 level, two-tailed test.

***Significantly different from zero at the 0.01 level, two-tailed test.

ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; ES = effect size; SD = standard deviation; TEMA-3 = Test of Early Mathematics Ability, Third Edition.

Table 7. Results by pre-K delegate urbanicity, ECLS-B and TEMA-3 post-tests

	Treatment			Control						
Pre-K delegate urbanicity	N	Ad-justed mean	Un-adjusted SD	N	Ad-justed mean	Un-adjusted SD	Impact estimate	SE of impact estimate	ES	*p*-value
ECLS-B post-test										
Urban	119	32.06	5.36	109	30.22	6.27	1.84	0.55	0.32***	0.001
Suburban	339	30.90	5.40	350	29.08	6.15	1.82	0.35	0.31***	0.000
Town/rural	195	30.08	5.58	201	28.49	6.32	1.59	0.41	0.27***	0.000
TEMA-3 post-test										
Urban	119	17.37	7.40	109	16.05	8.04	1.31	0.79	0.17*	0.096
Suburban	339	16.11	7.30	351	14.30	7.42	1.81	0.53	0.25***	0.001
Town/rural	163	15.06	7.44	175	13.40	7.24	1.67	0.49	0.23***	0.001

Source: Direct child assessment.

Note: Results adjusted for clustering using a two-level model: children within pre-K classrooms. We used the following baseline control variables: ECLS-B and TOPEL pre-tests, age, gender, race/ethnicity, language, and cohort. We calculated effect sizes using Hedges' g formula:

$$\frac{\widehat{mean}_T - \widehat{mean}_C}{SD_{pooled}}$$

where $\widehat{mean}_T$ equals the adjusted treatment group mean, $\widehat{mean}_C$ equals the adjusted control group mean, and SD_{pooled} is the pooled SD. We used the following formula to calculate the pooled SD:

$$SD_{pooled} = \sqrt{\frac{(N_T - 1)SD_T^2 + (N_C - 1)SD_C^2}{N_T + N_C - 2}}$$

where N_T equals the treatment group sample size, N_C equals the control group sample size, SD_T equals the unadjusted treatment group SD, and SD_C equals the unadjusted control group SD. To determine urbanicity, we used the Common Core of Data's urban-centric locale classification corresponding to the school district office location of each pre-K delegate.

*Significantly different from zero at the 0.1 level, two-tailed test.

***Significantly different from zero at the 0.01 level, two-tailed test.

ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; ES = effect size; SD = standard deviation; TEMA-3 = Test of Early Mathematics Ability, Third Edition.

Table 8. TEMA-3 effect sizes from current and recent studies, calculated consistently across studies

Study	Sample size	Unadjusted treatment post-test mean	Unadjusted control post-test mean	Post-test MD	Unadjusted control post-test SD	ITT ES	"Total" program impact
Study 2	670[a]	14.82	12.49	2.33	6.64	0.35	235
Study 3	240	14.87	11.83	3.04	6.81	0.45	107
Study 4	372	14.26	12.04	2.22	7.06	0.31	117
The current study	1256	15.96	14.49	1.47	7.51	0.20	246

Source: Direct child assessment; study author-reported results.

Note: We calculated effect sizes using the following formula:

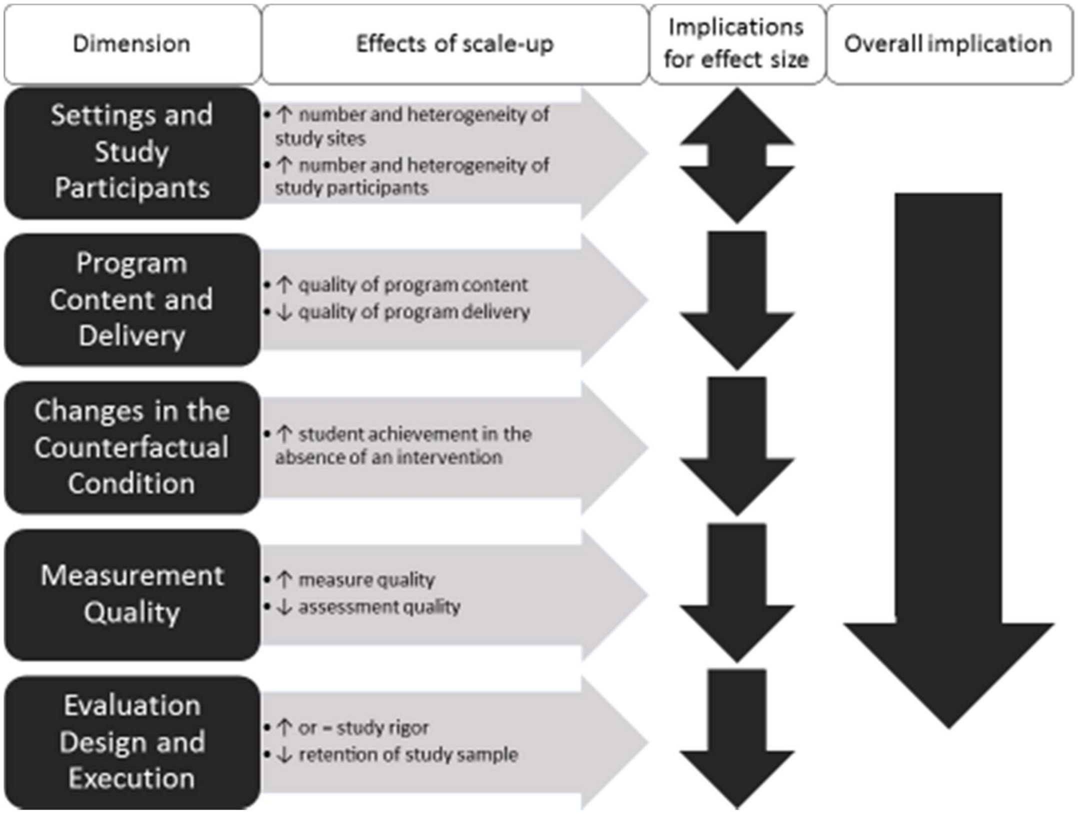
$$\frac{mean_T - mean_C}{SD_C}$$

where $mean_T$ equals the unadjusted treatment group mean, $mean_C$ equals the unadjusted control group mean, and SD_C is the unadjusted control group SD.
We calculated the "total" program impact by multiplying the ITT effect size by the sample size for each study.

[a] Sample size estimated from author-reported baseline sample size and overall attrition rate.

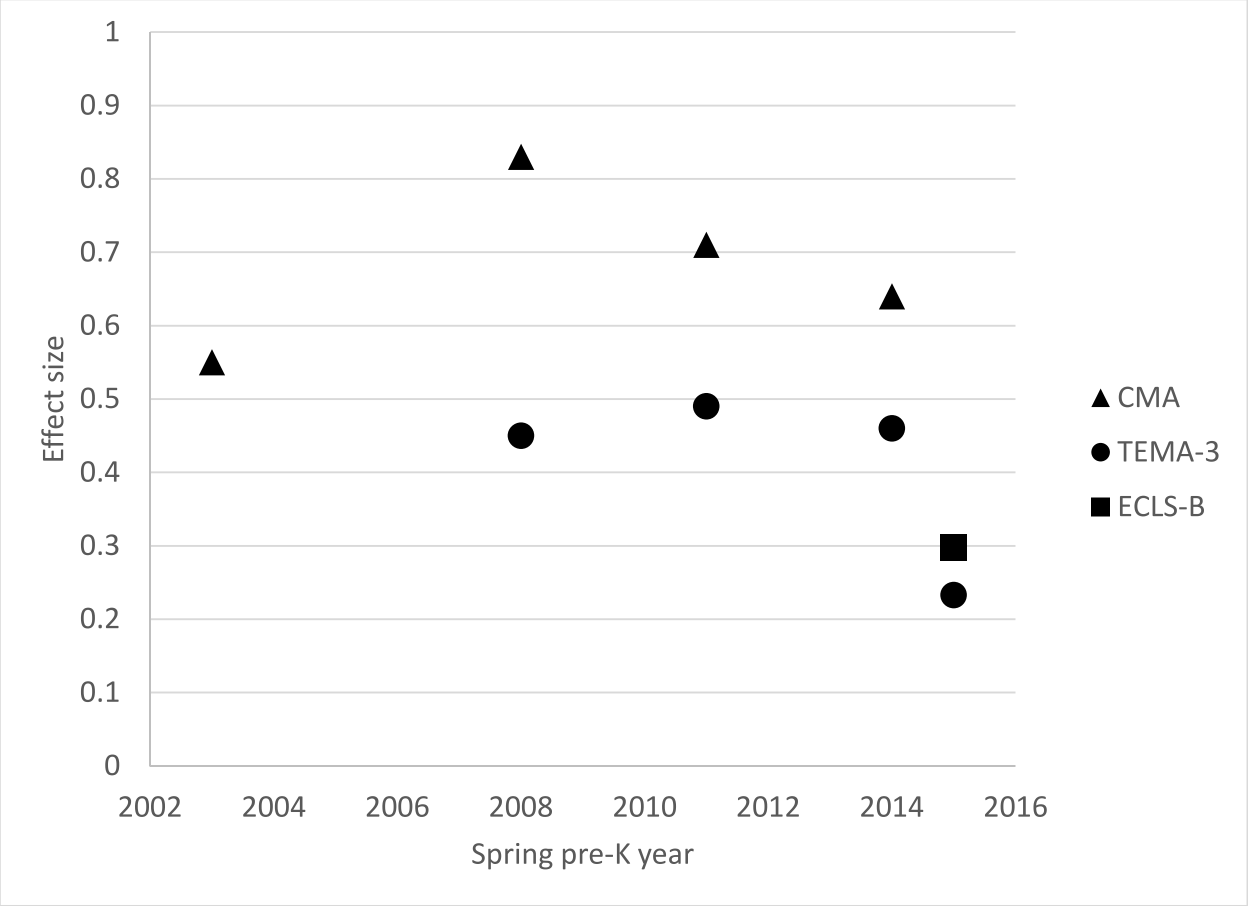
ES = effect size; ITT = intent-to-treat; MD = mean difference; SD = standard deviation; TEMA-3 = Test of Early Mathematics Ability, Third Edition.

Figure 1. Dimensions of scale-up



Dimension	Effects of scale-up	Implications for effect size	Overall implication
Settings and Study Participants	• ↑ number and heterogeneity of study sites • ↑ number and heterogeneity of study participants	↕	↓
Program Content and Delivery	• ↑ quality of program content • ↓ quality of program delivery	↓	
Changes in the Counterfactual Condition	• ↑ student achievement in the absence of an intervention	↓	
Measurement Quality	• ↑ measure quality • ↓ assessment quality	↓	
Evaluation Design and Execution	• ↑ or = study rigor • ↓ retention of study sample	↓	

Figure 2. Effect sizes from current and prior studies of *Pre-K Mathematics*, authors' preferred calculations



Source: Direct child assessment; study author-reported results.

Note: This figure reports study authors' preferred effect size calculations. For the current study, we report ECLS-B and TEMA-3 effect sizes as calculated in Table 4.

CMA = Child Math Assessment; ECLS-B = Early Childhood Longitudinal Study, Birth Cohort Mathematics Assessment; TEMA-3 = Test of Early Mathematics Ability, Third Edition.