Embedding Causal Research Designs in Pre-K Systems at Scale

*Rachel Abenavoli, Natalia Rojas, Rebecca Unterman,
Elise Cappella, Josh Wallack, and Pamela Morris*

Summary

In this article, Rachel Abenavoli, Natalia Rojas, Rebecca Unterman, Elise Cappella, Josh Wallack, and Pamela Morris argue that research-practice partnerships make it possible to rigorously study relevant policy questions in ways that would otherwise be infeasible.

Randomized controlled trials of small-scale programs have shown us that early childhood interventions can yield sizable benefits. But when we move from relatively small, tightly controlled studies to scaled-up initiatives, the results are often disappointing. Here the authors describe how their partnership with New York City's Department of Education, as the city rapidly rolled out its universal pre-K initiative, gave them opportunities to collect experimental and quasi-experimental evidence while placing a minimal burden on educators.

They argue that this type of research can answer the most pressing ECE questions, which are less about whether ECE can make a difference and more about the conditions under which early interventions are effective at scale. They offer three recommendations for researchers, policy makers, and practitioners who are considering partnership work: build a foundation of trust and openness; carefully consider whether rigorous causal research or descriptive research is the right choice in a given situation; and be flexible, seeking opportunities for rigorous research designs that may already be embedded in early childhood education systems.

Rachel Abenavoli is a research assistant professor of applied psychology at New York University's Steinhardt School of Culture, Education, and Human Development. Natalia Rojas is a postdoctoral research associate at NYU Langone's Center for Early Childhood Health and Development. Rebecca Unterman is a research associate at MDRC. Elise Cappella is a professor of applied psychology at NYU Steinhardt and vice dean for research. Josh Wallack is deputy chancellor for Early Childhood Education and Student Enrollment in the New York City Department of Education. Pamela Morris is a professor of applied psychology at NYU Steinhardt and a faculty affiliate of the NYU School of Global Public Health.

Rachel Abenavoli, Natalia Rojas, Rebecca Unterman, Elise Cappella, Josh Wallack, and Pamela Morris

Despite decades of research and substantial public investment in early childhood education (ECE), there is still a wide gap between education research and education practice. What happens in classrooms and school districts isn't always based on the latest research evidence, and research studies aren't always designed to solve key problems of practice or yield information with direct implications for practitioners or policy makers. Although tightly controlled research studies such as randomized controlled trials produce the strongest evidence of programs' efficacy, findings from these studies—though methodologically rigorous—aren't always relevant to fully scaled local systems with broader priorities, unique constraints, and large, diverse populations to serve.[1] Research-practice partnerships (RPPs) are a promising approach to make education research more relevant because they help align the focus of research and the needs and priorities of practitioners and policy makers.[2] They also provide unique opportunities to rigorously evaluate program models at scale and draw clear, causal conclusions about programs' effectiveness in local contexts.

This article describes the development of co-designed studies conducted in the context of a multiyear partnership between researchers at the Steinhardt School of Culture, Education, and Human Development at New York University (NYU) and early education leaders in New York City's Department of Education (DOE) after the city began an ambitious reform: an effort in just two short years to turn a preschool system that served about 19,000 children into one that served about 70,000. In the initial phase of our partnership, our teams developed capacity and infrastructure to monitor the rapid expansion of Pre-K for All, as the program is called. NYU brought existing research evidence and descriptive data to bear on the city's important policy questions, and the DOE shaped the emerging research agenda by identifying priorities, sharing information and data, and building internal capacity for collecting and using data. Building on this critical early work, regular communication across our teams, and a foundation of trust established in the early years of the partnership, we've developed a research agenda to test a core component of the system: the Pre-K for All professional learning (PL) model for supporting program quality. Our jointly designed research studies focus on the city's high-priority questions using a range of rigorous methods that yield causal, actionable evidence about PL and its effects on classroom processes and child outcomes. By working together to rigorously answer questions of high practical relevance as the programs are being implemented at scale, our partnership aims to generate the kind of information that is useful to policymakers and researchers in the city and beyond to best support the learning and development of young children through prekindergarten at scale.

Embedding Rigorous Evaluations in Systems at Scale

Prekindergarten, on average, can improve children's academic and social outcomes, at least in the short term.[3] But its effects vary as a function of program quality. Developmentally appropriate classroom curricula and teacher PL are hallmarks of high-quality prekindergarten programs, and program quality is associated with gains in children's school readiness skills.[4] As in other disciplines, randomized controlled trials meet the highest standards of scientific

rigor and provide the strongest evidence of effective approaches. Several tightly controlled evaluations have shown that specific curricula targeting math, literacy, or social-emotional skills, together with aligned support for teachers, lead to changes in teaching practice and children's skills in the targeted domains.[5] Using randomized trials to document programs' efficacy is a critical step in translating research to practice, but there's a large discrepancy between the magnitude of impacts produced by programs tested in relatively small, tightly-controlled studies and those produced by national, state, or district-wide programs implemented at scale in the real world.[6] This discrepancy may be due in large part to differences in program implementation and support (for example, less frequent or less intense coaching, more variable implementation in the classroom), either because programs are too costly for districts to fully implement across an entire system or because the initial developers lack the capacity to support large-scale dissemination of programs in their original form.[7] The gap between impacts produced in efficacy trials and those achieved at scale could also be due to differences in population (for example, differences among the children served, the teachers doing the implementing, or the coaches supporting them) or even to unintended negative consequences (for example, program quality that is poor or inequitably distributed at scale), especially if rapid scale-up outpaces the capacity to support dissemination.[8]

Given the disconnect between impacts documented in smaller efficacy trials and impacts produced at scale, we need research that can provide clear, causal, and actionable answers about how to create high-quality, large-scale prekindergarten programs. Randomization is the gold standard for

evaluating program impacts, but when evaluating fully scaled programs already in operation, the advantages associated with randomization must be balanced against practical considerations.[9] Experimental research designs are challenging to embed in large, complex systems, and randomization may be difficult for a number of reasons, including infeasibility, competing priorities, or concern about withholding potentially effective approaches from sites or children who need them. For example, certain research designs may not be possible or ethical in districts committed to offering prekindergarten to all children or PL opportunities to all teachers, or when program implementation is already well under way. Or districts may decide to target services to specific subgroups or allow individuals to opt in to programs when capacity and resources are limited, rather than mandate policies or specific curricula across a large and diverse system. These choices may represent the best decisions from a policy perspective, but they may limit research opportunities and preclude traditional randomized designs.

An RPP offers a unique opportunity to navigate the challenges of evaluating large prekindergarten systems, embed rigorous research designs at scale, and answer policy-relevant questions.[10] Factors that characterize RPPs—such as frequent and regular communication, a foundation of trust, and commitment to a shared vision and long-term collaboration—pave the way for research and policy partners to find research solutions that align with district priorities and constraints while maximizing the rigor with which research studies are designed and conducted.[11] In contrast to university-led research projects, where practice and policy partners may have only

a small role in planning, research-practice partnerships are well-positioned to recognize hidden opportunities for research that may otherwise be missed, to make small adjustments to existing processes that may be neutral for policy but beneficial from a research perspective, and to quickly identify challenges and their solutions.

Our multiyear partnership between NYU researchers and DOE leaders overseeing Pre-K for All shows how RPPs can maximize methodological rigor when evaluating programs and policies at scale. We've worked in a large, complex system with unique priorities and constraints to take advantage of existing opportunities for research (such as capitalizing on randomization that has already occurred) and embed new opportunities that increase our ability to draw clear, causal conclusions about program impacts (for example, by making slight adjustments to assignment processes). Our recent work seeks to answer policy-relevant causal questions about PL, a key component of the system expected to support program quality and therefore child outcomes. To generate useful information to advance practice and policy, we've sought methodological approaches that could answer our high-priority questions, were feasible to embed in the existing system, and met standards of scientific rigor that would permit causal inferences and contribute to early childhood education science.

History and Evolution of the Partnership

New York City's is the largest school district in the nation, serving 1.1 million students from prekindergarten through 12th grade. Pre-K for All represents a commitment to providing free, full-day, high-quality prekindergarten to every four-year-old in the city, and it is arguably one of the most swiftly and broadly deployed educational policy initiatives in the nation. In just two years, the DOE hired thousands of new teachers and more than tripled the number of children in preschool, from 19,000 students in 560 programs before the 2014 rollout to about 70,000 children per year in over 1,800 sites starting in 2015–16. To expand as rapidly as it did, Pre-K for All built on an existing (though decentralized) public and private early childhood system. Today, Pre-K for All serves about 60 percent of its students through programs in community-based organizations (called New York City Early Childhood Centers or NYCEECs) that contract with the DOE; the remaining programs are in district schools and DOE-created prekindergarten centers. All are held to the same quality standards, are integrated across data platforms for functions such as child enrollment and screening, are assessed similarly for program quality, and are offered the same PL opportunities and onsite coaching.

The NYU-DOE partnership began in the spring of 2014 when the city announced the Pre-K for All expansion. Soon after the announcement, NYU faculty gave a briefing on the state of ECE research to the city, and they were invited to collaborate with the DOE and the mayor's office. NYU then raised resources to capitalize on the first 12 months of implementation as an opportunity to monitor the expansion (for example, through descriptive analyses and data visualization; see below) and build a foundation for data collection, data-based decision-making, and research. At all phases, the NYU team provided technical assistance to the DOE—for example, by supporting the use and interpretation of data and

bringing prior research to bear on issues of policy and practice. At the same time, the DOE involved NYU in developing key decision-making processes and discussing the possibilities and constraints in bringing prekindergarten to scale. Over time, NYU's role shifted from key resource for the city's expansion to partner in conducting research.

In 2014–15, the first year of the partnership, the NYU team supported the DOE's efforts to launch Pre-K for All, helping the city take the pulse of the rapidly developing system. The DOE opened and licensed centers, hired teachers, and reached out to families to tell them about and help them sign up for Pre-K for All. Simultaneously, the DOE expanded its use of data for decision-making, built its own data team internally, and created a system of data collection and review that could support prekindergarten quality. The city also invested in a large-scale study of the program, led by the private research firm Westat, to learn from parents, administrators, DOE staff, and others about how the program was perceived and to get an early read on how children were faring across the city in the new system. NYU worked closely with Westat in this first year, helping to develop and administer tablet-based assessments of children's school readiness. The NYU team also created maps that combined key neighborhood characteristics, the location of Pre-K for All expansion sites, and child assessment data. Presenting this information visually allowed us to identify "hot spots," such as neighborhoods where children were most academically at risk. It also supported quality assurance and helped the DOE deploy its resources strategically. And the city was able to answer such questions as "Do Pre-K for All classrooms and sites vary widely in the extent to which they serve children at

greatest socio-demographic, academic, and behavioral risk?" and "Are the gains made by children during their Pre-K for All school year similar to or different from other large urban school systems?" This work together laid the foundation for the partnership.

In the partnership's second year, 2015–16, we made a coordinated effort to align Pre-K for All's quality infrastructure with research-based practice, using methods appropriate for a large and diverse system. NYU reviewed the research to help the DOE develop research-based quality standards and select evidence-based PL tracks or program models that were aligned to those quality standards (we discuss this process in more detail below). We also worked together to construct a data-based decision-making process to assign the sites to tracks and to allocate onsite coaching support (for example, how often sites would receive support and whether the coach would be an instructional coordinator, a social worker, or both). Our two teams communicated regularly via weekly phone calls, periodic in-person meetings, and frequent emails, and our partnership grew stronger.

In 2016–17, the partnership's third year, the NYU and DOE teams worked closely to develop observational and survey-based measures of teacher practice to guide implementation of the PL tracks. The NYU team also received funding from the Foundation for Child Development to provide information on the teacher workforce (see the article in this issue by Jacqueline Jones). As part of this work, NYU examined how teachers with different training and experience were distributed across the city and collected data on teachers' formal PL (training and coaching), informal PL (advice networks), and work climate (satisfaction,

support, and stress) to identify barriers and opportunities for bringing PL to classrooms.

Technical assistance, descriptive information, and data visualization were critical in the initial phase of the partnership, when the city's priorities included monitoring the Pre-K for All expansion and deciding how best to allocate support and resources. But as the city's chief concerns shifted from understanding the Pre-K for All landscape toward testing and strengthening components of the system, our research questions shifted from descriptive to causal. Given that PL is one of DOE's key supports to help Pre-K for All sites improve program quality, we jointly developed a set of research questions that focused on how PL affects teachers and children. We turn now to that work and our efforts to embed rigorous evaluations of PL in the system.

Embedding Rigor to Evaluate Professional Learning

As we said above, in its early phase the partnership developed a differentiated PL system that supports all prekindergarten teachers and leaders. Because needs vary among leaders and teachers and from site to site, the city decided to offer multiple PL options, called "tracks," each year. This approach better meets the needs of a large and diverse population and increases buy-in among program leaders and teachers, who have disparate training, experience, and interests. Each year, each site is assigned to one of several tracks, each with its own theme or focus. The first track to be introduced, in 2015–16, was NYC Pre-K Explore, which integrates the evidence-based math curriculum Building Blocks with interdisciplinary Units of Study that

the DOE developed to support children's higher-order thinking and problem-solving.[12] It evolved from discussions in the early phase of the partnership, when the DOE and NYU together reviewed evidence-based models to support children's learning. Three other tracks were introduced in 2016–17:

- NYC Pre-K Thrive, in which sites learn evidence-based strategies to support children's social-emotional development, grounded in research on the evidence-based family engagement intervention ParentCorps;[13]

- NYC Pre-K Create, an arts-based approach that integrates visual art, dance, theater, and music into instruction to promote learning and engagement; and

- NYC Pre-K Inspire, now called Teaching Team Learning Communities, which focuses on topics aligned to the district's quality standards (for example, creating a positive classroom culture, engaging children in meaningful activity) that support instructional goals for early childhood education.

Each year, Pre-K for All program leaders and teachers attend three to four full-day PL sessions tied to their tracks. All sites also get onsite support from an instructional coordinator and/or a social worker; the frequency of visits and the makeup of the support personnel (that is, instructional coordinator, social worker, or both) depends on the site's needs. Explore sites also receive coaching designed to support implementation of Building Blocks and

its integration with the Units of Study, and Create sites work with a teaching artist to support implementation in the classroom.

Today, much of the partnership's work focuses on which approaches are effective and for which outcomes. We're not trying to find out whether prekindergarten "works" relative to no prekindergarten. Research along those lines wouldn't tell the DOE what it needs to know to operate a system in which all four-year-olds are guaranteed a prekindergarten seat and all teachers participate in some form of PL. Instead, we're working to understand the implementation and impacts of different PL approaches with distinct theories of change and targeted outcomes for teachers and children. This focus can not only help the DOE refine and strengthen its system in New York, but also contribute to early childhood education more broadly.

In an ideal world, we might randomize prekindergarten sites to different models of PL, which would give us useful causal evidence about how the models affect the desired outcomes. But with the system already fully scaled and well under way, simple randomization of children or sites to PL tracks isn't always feasible or consistent with the DOE's other principles or priorities. The DOE has prioritized choice in Pre-K for All as a way to meet the needs of its large and diverse population of about 70,000 prekindergarten children in over 1,800 sites. To increase access and equity for all children, families can apply to prekindergarten anywhere in the city, no matter where they live. At the same time, site leaders' preferences are strongly weighted in the algorithm the DOE uses to place sites in PL tracks; this aligns with the DOE's stance that differentiating PL

by both needs and interests will maximize the benefits for teachers, and that having a choice in the PL assignment process will increase buy-in and implementation quality on the part of leaders and teachers. This is a sensible approach from a policy standpoint, but given the possibility of selection bias, it makes it harder to conduct research that permits strong causal inference. The strongest test of a program model is one that compares individuals who receive the program with individuals who don't receive the program *who are equivalent before the program begins* (and this is what randomization ensures); if different kinds of sites or families make different kinds of choices about what PL to receive or what prekindergarten program to attend, the assumption that teachers and children are equivalent from the start may not hold. Because traditional randomized designs are at odds with the city's commitment to choice, we've worked together to identify and/or embed randomization in the existing Pre-K for All choice processes, rather than impose a separate research design.

A Natural Experiment: The Explore Evaluation

As the DOE shifted in 2016 from expanding to strengthening the system, we had a series of conversations to map out new research efforts. Together, we decided to focus on how Pre-K for All's PL system affected outcomes at the setting level (for example, what changes do we see in teacher practices and classroom quality?) and to focus on the Explore PL track, the first track launched by the DOE. As we planned this work, we explored potential experimental and quasi-experimental designs (that is, designs in which researchers simulate a randomized controlled trial by identifying treatment

Rachel Abenavoli, Natalia Rojas, Rebecca Unterman, Elise Cappella, Josh Wallack, and Pamela Morris

and control groups, even though true randomization didn't occur) that would meet our dual goals of using rigorous methods and honoring the city's emphasis on choice. This emphasis on choice was (and continues to be) reflected in the DOE's PL track assignment process: before the 2016–17 school year, the DOE developed a systematic method for assigning prekindergarten sites to the four PL tracks based on criteria that included the tracks' limited capacity (that is, for certain tracks, there is a maximum number of sites that can be served each year), program leaders' preferences among the PL tracks, and the sites' needs. Sites' needs were determined by factors such as the proportion of children at the site who were from high-poverty neighborhoods or living in temporary housing. Given the selection factors (preference and need) that could contribute to track assignment, it was critical to carefully shape our evaluation to circumvent these factors at the design stage or address them through our sampling or analytic approach.

We first considered the gold standard for causal research—random assignment of sites to PL tracks. Because only a limited number of slots were available in three of the PL tracks (Explore, Create, and Thrive), the DOE's assignment algorithm was designed to randomly assign sites to tracks if certain conditions were met (for example, if the number of sites wanting to participate in a given track and meeting other criteria exceeded the number of slots remaining). We considered whether we could capitalize on this randomization for the Explore evaluation but determined that doing so wouldn't be possible in the 2016–17 school year. Although some randomization occurred, too few sites were randomized for this approach to give us an adequate sample size. In addition, although the process for

assigning sites to tracks was data-based and systematic, it was also multidimensional and complex. Thus it didn't easily lend itself to a research design used successfully in other evaluations where children above a clear cutoff receive an intervention (for example, children born before a certain date can enroll in prekindergarten) and are compared to similar children just below the cutoff who don't receive the intervention (this is called a *regression discontinuity* design).

Although these options weren't feasible, in the process of working closely together to develop our research plans, we discovered a hidden natural experiment of the sort frequently used in educational, economic, public health, and policy research. *Natural experiment* refers to an instance when two otherwise identical groups are affected differently by an event that is unrelated to either the treatment or the outcome and is outside of the researchers' control. Because a natural experiment mimics random assignment, it can answer casual questions while avoiding some of the practical and ethical challenges inherent in experimental research at scale.

The natural experiment we uncovered was the lucky result of a delay in decisions about funding the Explore PL track; the delay occurred due to factors outside the strict control of the DOE, NYU, and the programs themselves. When it came time to assign sites to PL tracks, the DOE didn't know whether funding for a new cohort of Explore sites would be available. So the department ran the site assignment process under two scenarios. Scenario A assumed funding for Explore was available, and sites could be assigned to Explore, Create, Thrive, or Inspire. Scenario B assumed no funding for Explore was available, and sites could be

Figure 1. Explore Natural Experiment as a Result of Funding Delay
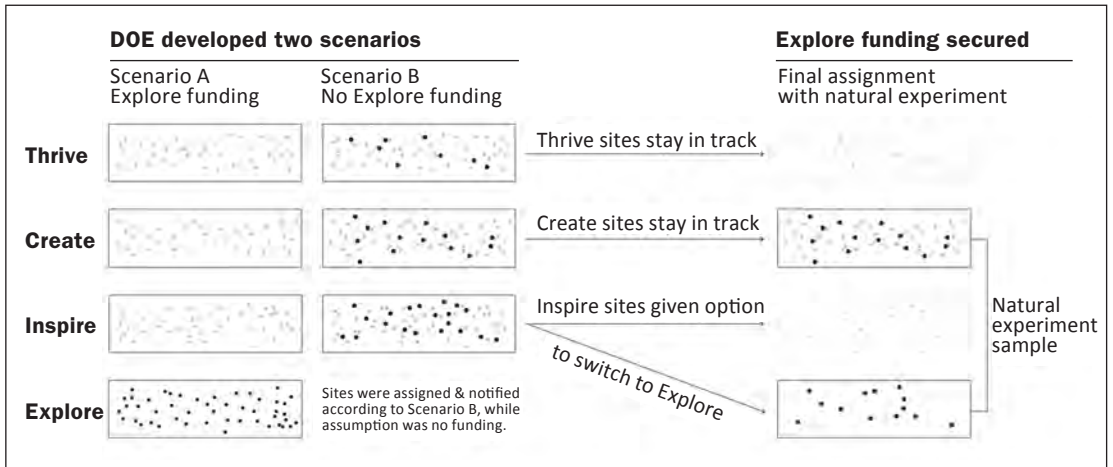
assigned only to Create, Thrive, or Inspire (see figure 1). Because funding for Explore seemed unlikely to materialize, the DOE notified sites of their track assignment under scenario B; that is, under the assumption that the Explore track wouldn't serve a new cohort of sites. Shortly thereafter, though, the DOE secured funding for an additional Explore cohort. Sites that had been assigned to Inspire under scenario B but would have been assigned to Explore in scenario A were offered the opportunity to switch to Explore; sites that had been assigned to the Create and Thrive tracks under scenario B weren't given the choice to switch because the DOE had committed to serve a certain number of program leaders and teachers in those tracks.

As figure 1 shows, the Explore track ultimately comprised sites that both had and hadn't been assigned to Explore under scenario A. We selected a subset of Explore sites as our study sample. The treatment group in our natural experiment included sites that would have been assigned to Explore in scenario A and, in the end, were assigned to receive it after all. Our comparison group comprised sites that

would have been assigned to Explore according to scenario A but were placed in a different track. This comparison group allowed us to examine what might have happened to sites in our Explore (treatment) group, on average, if they had been assigned to the other tracks. In our research, we use these inadvertent treatment and control groups to estimate the effects of Explore training and coaching on classroom quality and teachers' math instructional practices.

We were able to identify and take advantage of this natural experiment only because of the relationship we had established across our teams, which included open communication and frequent contact. Had the NYU team members been external researchers, we probably wouldn't have learned about the chain of events that led to the natural experiment; had the DOE not been willing to partner on this research and share details about the process—some of which may not have seemed relevant at first—we wouldn't have recognized the opportunity that the funding delay gave us to learn about Explore implementation and impacts. We expect that this is the case in other contexts as well. Natural experiments may not be as rare as

we might think, but they may go unnoticed when external researchers aren't in the know about such events and policymakers aren't on the lookout for such opportunities. RPPs are a way to avoid missed opportunities to rigorously estimate effects in the absence of a traditional randomized controlled trial.

Before our main analyses testing Explore's effects on program quality and teacher practice, we conducted preliminary analyses to ensure that the natural experiment was a valid approach for evaluating Explore. First, we examined whether treatment and control sites were equivalent at the beginning of the school year, as would be expected in a randomized controlled trial. Documenting equivalence would give us confidence that any differences we observed at the end of the year were due to the treatment (that is, to Explore PL) and not to pre-existing differences between sites. This was important to investigate because factors other than the natural experiment (for example, site leaders' second choice among the PL tracks) contributed to the final track assignments, given the way sites were reassigned after the DOE secured Explore funding. We tested differences between our treatment and control groups across a range of child- and site-level characteristics and found a few significant and/or marginal differences between Explore and non-Explore sites; for example, control group sites had higher classroom quality scores on a widely used rating scale. In all cases, the differences we observed had the potential to bias estimates of treatment impact downward rather than upward. But they still raised the possibility that observed differences in teachers' practices in the spring could reflect pre-existing differences rather than participation in Explore.

Following our primary analyses testing Explore's effects, we also ran supplemental analyses to examine whether our findings were consistent across different subsamples and under different analytic assumptions. For example, we tested whether our findings were similar across the full sample, among district schools only, and among sites with the same PL preferences (for example, sites that listed Explore as their first choice and Create as their second choice) since this ultimately influenced their track assignment. Restricting the sample in these ways allowed us to estimate impacts among treatment and control sites we knew to be equivalent on key characteristics at the start of the year. This strengthened our confidence that any positive impacts we observed at the end of the year were due to participation in Explore and not pre-existing differences, especially when different analytic approaches produced results that were similar in pattern, magnitude, and significance.

Using Existing Randomization: Application Lotteries

We've just seen that the close partnership and regular communication between NYU and the DOE let us capitalize on a natural experiment to answer a policy-relevant question when the opportunity presented itself and to generate critical information about the Explore PL track. But new research questions emerged as the Pre-K for All system continued to expand and evolve over time. As an RPP with the mission of using data and research to strengthen the system broadly, we were well-positioned to expand our focus and answer new questions. For example, how do the other PL tracks affect children's outcomes? What are the effects among new cohorts of sites and children?

Figure 2. Lotteries in Three Hypothetical Sites
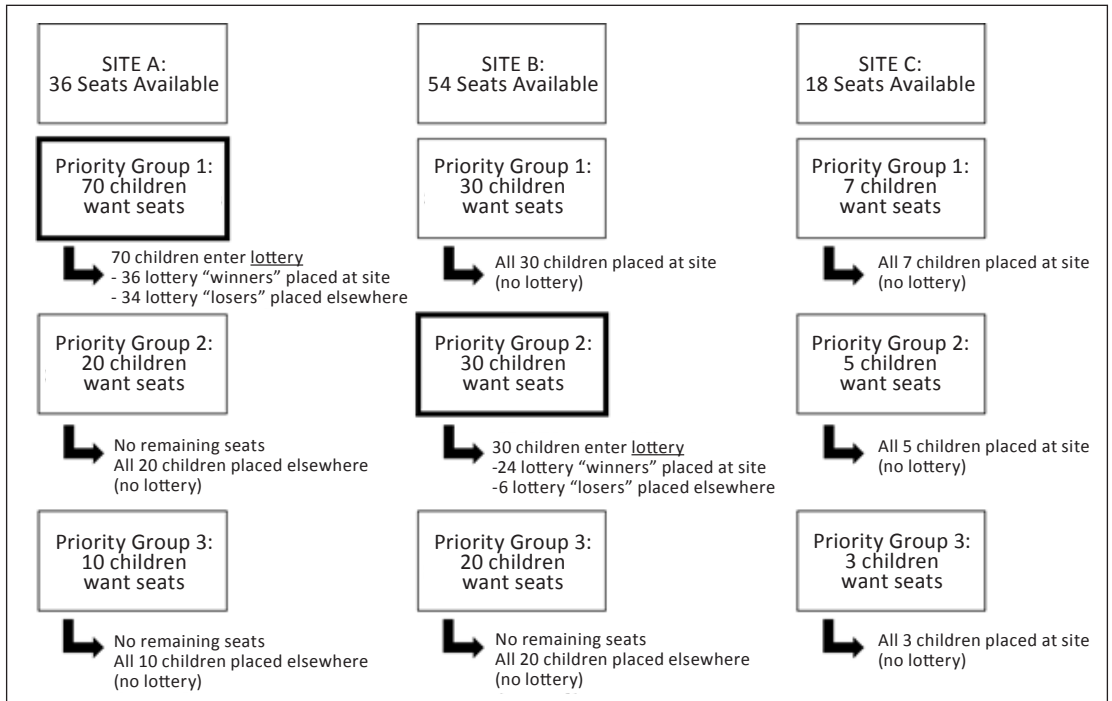
In the absence of natural experiments that could help us answer these new questions, we have considered other designs for understanding how PL affects teacher practice and child outcomes. One approach that's been used in New York at the high school level and increasingly in other settings at the prekindergarten level takes advantage of the fact that children are randomized to sites via lotteries that occur when a site receives more applications than seats; we call these sites *oversubscribed*.[14] In such cases, some children are randomly assigned to the oversubscribed site ("lottery winners") and other children are assigned to a site lower on their preference list ("lottery losers"). This lets researchers estimate both the effect of winning a lottery in and of itself (known as *intent-to-treat effect*) and the effect of winning a lottery and then enrolling in a given site (known as *local average treatment effect*).[15]

In New York, parents can apply to up to 12 Pre-K for All sites for their children, in order of preference. At the same time, the DOE places each child in a "priority group" for each site, which is based on criteria such as living in a certain zone or district, whether the student has a sibling at the site, and whether the child was a student at that site as a three-year-old (see figure 2). To understand whether we could use Pre-K for All application lotteries to answer our research questions about PL, we used a prior year of Pre-K for All application and enrollment data to examine issues of internal validity (the extent to which the child-level lottery design would allow us to establish the causal effect of PL) and generalizability (the extent to which our findings would apply to Pre-K for All children and sites outside our lottery sample). First, we identified the number and distribution of lotteries that occurred through the Pre-K for All

application process. Building on previous work that used lotteries to estimate the impacts of the city's Small High Schools of Choice, we developed an algorithm to identify sites where lotteries occurred, as well as the individual children who won or lost those lotteries.[16] To ensure that we accurately identified lotteries and lottery participants during this process, NYU researchers met several times with DOE partners to discuss the application process and nuances in the data. We then confirmed that the assignment process produced enough lotteries for each PL track to give us adequate sample size to test the PL tracks. We found that the lotteries were well distributed across the city and across different types of sites, indicating that lottery sites represented the full range of geographic and setting type diversity within the Pre-K for All system.

We also examined whether lotteries produced treatment and control groups (lottery winners and lottery losers, respectively) that were similar at the start of the school year. Because only limited demographic data were available on prekindergarten children and their families at the time of application, we used information about children's application choices themselves (for example, the number of choices listed, the proportion of choices in children's home districts, and the type and quality of children's top choices) as baseline characteristics, as well as demographic data that were available once children enrolled in Pre-K for All (but that weren't available for those who didn't enroll). Across these analyses, we found that the choice patterns and demographic characteristics of lottery winners and losers were generally well-balanced. This indicates that the lottery process did in fact create groups of treatment children and control

children who were similar at the start of prekindergarten, a necessary condition for drawing causal inferences about the effects of PL.

To better understand the *treatment contrast* (that is, the difference between what treatment children experience and control children experience), we also examined the extent to which lottery winners and losers complied with their lottery assignments and the effect of winning a lottery on children's prekindergarten experiences. As expected, we found that winning a lottery for a site in a given track was indeed associated with greater exposure to that track during the prekindergarten year (for example, winning a lottery for an Explore site increased children's exposure to Explore). However, analyses also revealed that winning a lottery was associated with exposure to site characteristics other than PL (for example, higher program quality), making it difficult to isolate the effects of PL under this design. NYU, the DOE, and a team of methodological experts then considered multiple options for addressing the fact that lottery winners and lottery losers attended sites that differed in ways other than PL. For example, we considered using complementary quasi-experimental methods to test each track relative to a matched (but non-randomized) control group comprising only children in Inspire sites, and using secondary, descriptive analyses to examine how much of the treatment effect was attributable to PL versus other site characteristics. We also considered and ultimately switched to a different design altogether—a site-level randomized design—when we learned that sites would be assigned to PL track via the DOE's track assignment process. In our case, this design turned out to be better than the child-level lottery design because

it allowed us to experimentally isolate the effects of PL (our primary research question) and produce the clearest and most relevant evidence for NYC DOE.

Creating Randomization in Track Assignment

We've described two examples of how events or processes occurring in the education system can be used to conduct rigorous evaluations at scale: a natural experiment involving site-level assignment that occurred at a single point in time, and child-level randomization that occurs each year during the application process. In both examples, the existing system is taken as a given; that is, we capitalized on randomization that occurs as a by-product of the system and didn't manipulate the system for research purposes. In contrast, a third approach we've explored involves tweaking the system to create new opportunities for rigorous research.

Recall that Pre-K for All sites are assigned to PL tracks through a complex algorithm. Although the algorithm is revisited each year, assignment has always relied on program leaders' rank-ordered preferences among the PL tracks, site need, and each track's capacity. Our NYU and DOE teams have worked together to explore whether and how this PL assignment process might be adjusted to ensure more randomization to enable rigorous research in the future.

In our discussions, we considered the advantages of randomization for understanding the causal impacts of the different tracks and producing clear, actionable findings the city could use to strengthen and refine its PL model. The DOE is interested in carrying out more research on PL tracks, and it recognizes the
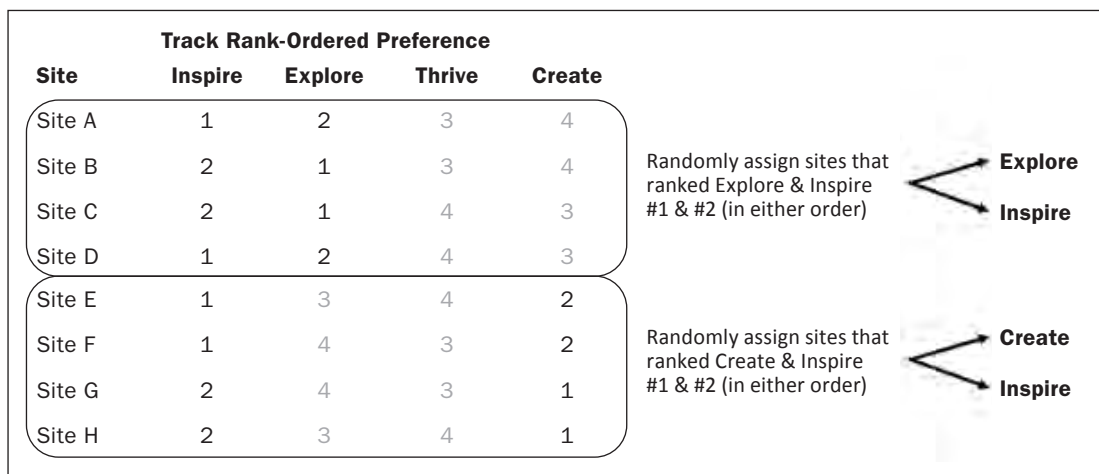
value of maximizing the rigor of that research. But it has also underscored a rationale for prioritizing program leaders' preferences in track assignment—they see this approach as critical both for meeting the sites' individual needs and interests and for building buy-in among leaders and teachers. Indeed, the DOE's early data on teachers' attendance at PL sessions indicated high attendance among programs that received their first or second choice, with a notable drop-off in attendance for programs who received their third-choice track. Keeping in mind the two goals of honoring program leader preference and conducting high-quality research as our starting points, our partnership has searched for ways to embed randomization in the DOE's process for assigning sites to PL tracks in a way that still prioritizes choice.

Figure 3 illustrates one approach we came up with, simplified for illustration purposes (for example, we haven't included various preceding steps or exclusion criteria). Briefly, we jointly developed a process in which subsets of sites could be stratified by similar first and second preferences among tracks and then randomized. For example, sites that listed Explore and Inspire as first and second choices (in either order) would be grouped together and then randomized to either Explore or Inspire. Through this process, all the sites would receive either their first or second choice. At the same time, site-level randomization would make it possible to evaluate the causal effects of one track relative to another among sites with similar preferences (and other characteristics). Thus we can embed rigor and opportunities for research while still honoring program leaders' preferences.

As in the Explore evaluation we described above, we wouldn't have thought of this

Figure 3. Randomizing Hypothetical Sites Based on Preference Subgroups

	Track Rank-Ordered Preference			
Site	**Inspire**	**Explore**	**Thrive**	**Create**
Site A	1	2	3	4
Site B	2	1	3	4
Site C	2	1	4	3
Site D	1	2	4	3
Site E	1	3	4	2
Site F	1	4	3	2
Site G	2	4	3	1
Site H	2	3	4	1

Randomly assign sites that ranked Explore & Inspire #1 & #2 (in either order) → **Explore** / **Inspire**

Randomly assign sites that ranked Create & Inspire #1 & #2 (in either order) → **Create** / **Inspire**

approach outside the context of an RPP built on a history of joint efforts and a foundation of trust and mutual respect for each other's perspectives, expertise, and commitment. Finding this potential solution required flexibility on the part of both researchers and policy partners as we worked together toward the shared goal of embedding research that generates the strongest evidence for understanding and strengthening the PL system.

Challenges

We've shown how the NYU-DOE partnership navigated a complex system to find or create opportunities to conduct rigorous and relevant research. But these examples also reveal challenges. Any research project involves tradeoffs, and the designs we've used in our work have limitations. In the case of the Explore natural experiment, we found that despite randomness in track assignment due to funding delays, the treatment and control groups weren't fully equivalent on all site-level characteristics. In our broader evaluation of the Pre-K for All PL system, we've faced other challenges. For example, although the lotteries that are created during the application process ensure that children who win lotteries (treatment participants) and lose lotteries (control participants) are equivalent in expectation, the sites they attend are not. This means that the treatment contrast we can estimate using this experimental approach isn't as precise as we might like, given that the sites that lottery winners attend differ from the sites that lottery losers attend with regard not only to PL tracks but to other characteristics as well. These differences may contribute to differences in children's experiences and outcomes, which required us to consider complementary methods and alternative designs to isolate the effect of PL.

The methodological issues we've encountered underscore a few important points. For one thing, conducting rigorous research in fully scaled systems is messy. We've found that the advantages of capitalizing on rigor that already exists in such a system outweigh the disadvantages. But it's not always a perfect solution. It's critical to conduct an extensive set of analyses to examine the validity and limitations of the intended research designs, and then to revisit and revise the designs or analytic approach

on the basis of these findings. Using multiple analytic approaches to answer the same research question may also be a useful way to allay some of the methodological challenges of evaluating fully scaled programs.

Another problem is that the research process may be slower than the evolution of the system, program, or population we seek to study, a challenge other partnerships have also experienced (for example, see the article in this issue by Christina Weiland and Jason Sachs). The delay between planning a study and executing it can challenge both researchers and their practice and policy partners. Because local education programs or initiatives are constantly evolving, research designs and measurement protocols that may be perfectly appropriate at a given time may not be appropriate a year later because components of the system have changed. Indeed, these sorts of changes are expected and desired by leaders who work to improve the system's implementation. From the policy partners' perspective, multiyear research projects may cause district leaders to feel locked in to certain decisions or restricted in the programmatic changes they could make to strengthen the system.

We've encountered some of these challenges of studying an evolving system in New York City, where the DOE invests significant time and resources to improve the quality of Pre-K for All—including efforts to modify and refine the PL model that is a focus of our joint research. On one hand, this means that the system is continually improving and innovating, which ideally leads to better outcomes for children. On the other hand, our research may be too slow at times to contribute to these changes, and the right time to freeze and evaluate a constantly evolving system isn't always clear. The

communication structures in our RPP and a foundation of trust built over the years have made it possible for the DOE to raise these kinds of concerns with NYU early and work toward mutually agreeable solutions. Flexibility has been key in this process: thanks to flexibility in our funding and flexibility in our partnership, we've adjusted aspects of our research plans while staying true to the key aim of evaluating the impacts of different PL models on teachers and children.

Another challenge of partnership work at scale is navigating a large district central office and a large university. To ensure shared understanding and alignment across partners, we've relied on frequent communication through multiple channels with the many people who have a stake in our work. We've also thought about how and when to loop in researchers and/or policy partners during meetings and planning processes versus when to move the work forward internally in our respective organizations. For example, both NYU and DOE need to be involved when developing research questions or discussing possible research approaches. But examining the validity or limitations of particular methodological approaches likely doesn't require as much DOE involvement. Likewise, rather than involve NYU researchers in every planning meeting before a new initiative is rolled out, in some cases the DOE brings in researchers only at key decision points, toward the end of a planning process, or when there may be implications for research.

Successes

Our partnership's work has a number of implications for practice and policy in New York City, for early childhood education more broadly, and for the way we use and conduct research within and outside our

partnership. Most importantly, our work supports the DOE's efforts to refine and strengthen Pre-K for All. Over the course of our partnership, NYU has helped the DOE by sharing emerging trends in research, lending expertise about methods and measurement tools, and telling the DOE about other districts undertaking similar efforts or facing similar problems. The DOE has identified priorities, shaped our research agenda, provided opportunities for rigorous research, and navigated potential barriers to our joint work. By thinking and planning together, particularly around PL and program quality standards, we developed the framework for monitoring and supporting quality in prekindergarten sites across the city. As findings emerge from our evaluations of the PL system, these too will guide further adaptations of the Pre-K for All system—including both incremental improvements such as strengthening onsite implementation support and larger changes such as discontinuing tracks that aren't shown to be effective.[18] By identifying components of the system that are working well, as well as areas that could be improved, our co-conducted research provides concrete information for decision-making and helps the DOE secure additional funding to support effective program models. Indeed, promising findings from the first year of the Explore natural experiment provided initial support for continuing to invest in that track. Understanding whether certain approaches are more effective for different types of sites or children will be especially important as the DOE seeks to refine differentiation of PL within the system, including the ways that PL might be sequenced over multiple years.

Our partnership's work has also strengthened the quality and relevance of our research, and therefore enhanced our contributions to education science broadly. New York is unique in some important ways, but findings from our work are generalizable to other large urban districts. We expect our work to advance the scholarship on PL in early child education by focusing on implementation and impacts in fully scaled real-world systems. As we've said, most of the rigorous evidence about whether PL models or curricula are effective comes from smaller, tightly controlled trials where program developers typically had a hand in program implementation and/or evaluation. By identifying programs that are effective at scale under real-world conditions across a diverse population, our work will produce evidence that's more relevant to the realities faced by cities, states, and districts and thus more likely to support practice and policy.

Our work has also increased the capacity of NYU and the DOE to sustain research-practice partnerships. By developing structures for regular communication and feedback, engaging partners at different levels of the organization, and establishing a successful track record, we've developed a culture and shared language around rigorous research, evidence, and data. That's allowed us to jointly identify new research opportunities and new ways to use data or prior research to guide decision-making.

Our experiences in the partnership have influenced our work outside the partnership as well. As policy makers, we've increasingly emphasized data and research in other areas of decision-making. As researchers, we've become more flexible and creative with the research tools at our disposal and learned to find opportunities for rigor even when we don't have strict control over randomization. In addition, in the examples we've described here, we examined the

extent to which the research designs were appropriate for answering our questions of interest. We also identified limitations in our designs and came up with ways to overcome them. These issues are particularly salient in our work evaluating Pre-K for All, where we as researchers have less control over the assignment processes underlying our research designs. But carefully testing the assumptions and validity of a given approach is no less important in other research. The same can be said about whether a sample is representative of the population being studied and whether results can be generalized to other contexts. Given that the DOE needs to understand how PL is working across the system, we've had to carefully consider how findings from the Explore natural experiment sample and the lottery subsample might or might not be applicable to other Pre-K for All sites and children. Paying more attention to these issues has carried over into our work outside the partnership.

Recommendations for Researchers and Policymakers

RPPs are uniquely positioned to capitalize on existing opportunities and/or negotiate small tweaks in existing systems to strengthen the rigor of research designs. In our experience, this is made possible by a sense of trust built over the years that characterizes effective and sustained partnerships. Open communication channels and flexibility on the part of both researchers and policy partners have also been critical. Below we offer three recommendations for researchers, policymakers, and practitioners who are considering partnership work to advance the science and practice of early childhood education.

Build a Foundation of Trust and Openness to Research.

The current phase of our research is possible only because of the joint work that preceded it and the relationships we cultivated through years of close collaboration. In our partnership, NYU has remained committed to supporting the DOE's efforts across a range of initiatives, and the DOE has remained open to and invested in high-quality research as a means to strengthen Pre-K for All. From its earliest stages, NYU and the DOE made this partnership a high priority, followed through with timely and high-quality work, and dedicated staff and funding to support the effort. We've found that shared understanding of—and responsiveness to—each partner's goals, concerns, and expertise are critical to the partnership's health. A foundation of trust allows partners to be open and candid, which helps in building a relevant and rigorous research agenda. Identifying mutually agreeable research aims and designs comes more easily when policymakers are transparent about their goals and priorities beyond research and when researchers are honest about the advantages, limitations, and potential burdens on the system of different designs.

We encourage anyone looking to establish or sustain an RPP to engage members at multiple levels of each institution, communicate regularly, and find opportunities to meet in person. Engaging a range of people with varied roles and expertise helps to build a culture of using research to improve practice, helps sustain the partnership as people transition in and out, and brings new opportunities to light that might be missed if collaboration were confined to a select few. Establishing project-specific meetings has helped us focus on a particular task or

issue, while standing check-ins allow partners to share general updates, raise issues or questions, or tee up new ideas. This structure ensures that we make steady progress on planned work while also letting new research opportunities come to light as we keep each other in the loop about other efforts outside the partnership. These factors helped us uncover and understand the Explore natural experiment, for example, and identify ways to embed randomization in the process for assigning prekindergarten sites to tracks. Regular check-ins, with many members of our partnership at the table, have let us identify and capitalize on new opportunities for research or research-based decision-making as they emerge, or shape our research agenda in ways that better align with new priorities or upcoming initiatives.

Consider the Right Time for Causal versus Descriptive Research

Although rigorous evaluation is often critical to guide practice and advance ECE scholarship, descriptive research can also be useful for RPPs. As we said earlier, our partnership's initial research efforts were primarily descriptive, and we shifted to causal methods only when our jointly developed research agenda changed its focus to the PL system and its impacts. Our perspective is that trustworthy, well-sampled descriptive data are invaluable to senior leaders as a foundation to evaluate whether new policy initiatives are on the right track. Moreover, we expect that senior leaders benefit from gaining fluency with quantitative data about children's school readiness and program quality as a necessary first step toward more complex questions that might involve more sophisticated analytic approaches.

A few factors contributed to our shift from descriptive to causal research. First, the city's

priorities changed from monitoring the system during a period of rapid expansion to strengthening quality within it. Second, after a few years of Pre-K for All, the DOE and NYU agreed that certain components of the system were ready for evaluation. We first focused on impacts on program quality and then expanded our focus to child outcomes once the system stabilized to some degree after a period of rapid expansion and development. Third, by this point, our partnership was strong enough—thanks to trust built during the first phase—to engage in evaluations with higher stakes than the descriptive work we first took on. Other factors might come into play in other settings, such as program scale, resource or other constraints, the speed with which results are needed to make decisions, the structure of the system or program under investigation, or the nature of the process being studied. It should be clear that we don't recommend maximizing rigor at all costs. Researchers and policy makers should carefully consider when and where causal versus quasi-experimental versus descriptive research is most appropriate given the research questions and context.

Be Flexible

With the examples in this article, our intention is to encourage research and policy partners to think creatively together about:

1. whether rigor that can be used for research purposes may already exist in a system and;

2. whether there are ways to add rigor while staying true to other priorities and operating within existing constraints.

Indeed, certain programmatic constraints such as limited funding or capacity present unique opportunities for rigor, and these can be used to the advantage of research and policy partners rather than being seen as barriers. In this article, we discussed three approaches to rigorous research that fit our needs and constraints in New York City—capitalizing on a natural experiment, using existing randomization of children to oversubscribed prekindergarten programs, and creating new opportunities for rigorous research by incorporating randomization into the way sites are assigned to PL tracks. We encourage research and policy partners to consider these types of designs, as well as other approaches. For example, waitlist designs (when an intervention is withheld from a control group for a certain period of time but is offered to that group once the research is complete) may be suitable when a new program or initiative can't be rolled out instantly; randomization may be possible when an intervention has limited capacity to serve prekindergarten teachers or students; regression discontinuity designs may work well when clear cutoffs are used to make decisions about how to allocate resources or assign children or sites to interventions. And as we've said, descriptive work may be preferable depending on the focus of the research question, the stage of the program being studied, or the stage and strength of the partnership itself. Finding the optimal design requires researchers to be willing to deviate when necessary from randomized controlled trials and explore other, more feasible approaches that still maintain high standards of rigor, including quasi-experimental designs when randomization isn't possible in any form.

It is also worth noting that our partnership was established to support the monitoring and strengthening of the Pre-K for All system broadly, rather than to answer a specific question or evaluate a specific component of the system. That has given us the flexibility to revisit and revise research aims and successfully adapt to evolving priorities and emerging questions, setting our approach apart from more traditional university-driven research. For example, we shifted from descriptive work to evaluation of the PL system when the city needed answers about how to improve quality, and we adapted our early plans to evaluate impacts on child outcomes to evaluate impacts on program quality after recognizing that program-level results would be more helpful to the DOE in the first few years of the PL system's implementation. The guiding principle for our multiyear collaboration thus far has been to put science to work, and our process has supported our broad goal of a sustained partnership that facilitates bidirectional influence of rigorous research and policy/practice in ways that are useful to the city first and to education science generally second.

Our work shows that it's possible through partnerships to build studies that can provide strong causal evidence that's directly relevant to policymakers and meets the standards of rigor that are necessary for scholarship, even when constraints or competing priorities in large systems make it impractical or impossible to use randomized controlled trials. With flexibility and persistence, researchers and policy partners are well positioned to produce the kind of evidence that advances understanding of key issues in ECE, directly supports programming and decision-making on the ground, and maximizes benefits for prekindergarten leaders, teachers, and children.

Endnotes

1. John Q. Easton, "IES: Promises and Challenges" (lecture, Annual Meeting of the American Educational Research Association, Philadelphia, PA, April 5, 2014); C. Cybele Raver, "Targeting Self-Regulation through Intervention: Lessons from RCTs" (lecture, Society of Research on Educational Effectiveness Fall Conference, Washington, DC, September 26, 2013).

2. Vivian Tseng, John Q. Easton, and Lauren H. Supplee, "Research-Practice Partnerships: Building Two-Way Streets of Engagement," *Social Policy Report* 30, no. 4 (2017): 3–15, https://doi.org/10.1002/j.2379-3988.2017.tb00089.x.

3. Gregory Camilli et al., "Meta-Analysis of the Effects of Early Education Interventions on Cognitive and Social Development," *Teachers College Record* 112 (2010): 579–620; Deborah A. Phillips et al., *The Current State of Scientific Knowledge on Pre-Kindergarten Effects* (Washington, DC: Brookings Institution, 2017); Hirozaku Yoshikawa et al., *Investing in Our Future: The Evidence Base on Preschool Education* (Washington, DC: Society for Research in Child Development, 2013).

4. Eric A Hanushek, "The Economic Value of Higher Teacher Quality," *Economics of Education Review* 30 (2011): 466–79, https://doi.org/10.1016/j.econedurev.2010.12.006; Andrew J. Mashburn et al., "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills," *Child Development* 79 (2008): 732–49, https://doi.org/10.1111/j.1467-8624.2008.01154.x; Christina Weiland et al., "Launching Preschool 2.0: A Road Map to High-Quality Public Programs at Scale," *Behavioral Science & Policy* 2, no. 1 (2016): 37–46; Steven W. Barnett, "Effectiveness of Early Educational Intervention," Science 333 (2011): 975–8, https://doi.org/10.1126/science.1204534; Margaret Burchinal et al., "Predicting Child Outcomes at the End of Kindergarten from the Quality of Pre-Kindergarten Teacher–Child Interactions and Instruction," *Applied Development Science* 12 (2008): 140–53, https://doi.org/10.1080/10888690802199418; Tran D. Keys et al., "Preschool Center Quality and School Readiness: Quality Effects and Variation by Demographic and Child Characteristics," Child Development 84 (2013): 1171–90, https://doi.org/10.1111/cdev.12048; Yoshikawa et al., *Evidence Base*.

5. Karen L. Bierman et al., "Executive Functions and School Readiness Intervention: Impact, Moderation, and Mediation in the Head Start REDI Program," *Development and Psychopathology* 20 (2008): 821–43, https://doi.org/10.1017/S0954579408000394; Douglas H. Clements et al., "Mathematics Learned by Young Children in an Intervention Based on Learning Trajectories: A Large-Scale Cluster Randomized Trial," *Journal for Research in Mathematics Education* 42 (2011): 127–66, https://doi.org/10.5951/jresematheduc.42.2.0127; Alice Klein et al., "Effects of a Pre-Kindergarten Mathematics Intervention: A Randomized Experiment," *Journal of Research on Educational Effectiveness* 1 (2008): 155–78, https://doi.org/10.1080/19345740802114533; Christopher J. Lonigan et al., "Promoting the Development of Preschool Children's Emergent Literacy Skills: A Randomized Evaluation of a Literacy-Focused Curriculum and Two Professional Development Models," *Reading and Writing* 24 (2011): 305–37, https://doi.org/10.1007/s11145-009-9214-6; Pamela Morris et al., *Impact Findings from the Head Start CARES Demonstration: National Evaluation of Three Approaches to Improving Preschoolers' Social and Emotional Competence* (New York: MDRC, 2013); C. Cybele Raver et al., "Improving Preschool Classroom Processes: Preliminary Findings from a Randomized Trial Implemented in Head Start Settings," *Early Childhood Research Quarterly* 23 (2008): 10–26, https://doi.org/10.1016/j.ecresq.2007.09.001; C. Cybele Raver et al., "Targeting Children's Behavior Problems in Preschool Classrooms: A Cluster-Randomized Controlled Trial," *Journal of Consulting and Clinical Psychology* 77 (2009): 302–16, https://doi.org/10.1037/a0015302; Barbara A. Wasik and Mary Alice Bond, "Beyond the Pages of a Book: Interactive Book Reading and Language Development in Preschool Classrooms," *Journal of Educational Psychology* 93 (2001): 243–50, https://doi.org/10.1037%2F0022-0663.93.2.243; Carolyn M. Webster-Stratton, Jamila Reid, and Mike Stoolmiller, "Preventing Conduct Problems and Improving School Readiness: Evaluation of the Incredible Years Teacher and Child Training Programs in High-Risk Schools," *Journal of Child Psychology and Psychiatry* 49 (2008): 471–88, https://doi.org/10.1111/j.1469-7610.2007.01861.x.

6. Greg J. Duncan and Katherine Magnuson, "Investing in Preschool Programs," *Journal of Economic Perspectives* 27 (2013): 109–32, https://doi.org/10.1257/jep.27.2.109; Robert C. Pianta et al., "The Effects of Preschool Education: What We Know, How Public Policy Is or Is Not Aligned with the Evidence Base, and What We Need to Know," *Psychological Science in the Public Interest* 10 (2009): 49–88, https://doi.org/10.1177/1529100610381908.

7. Bridget K. Hamre, Ann Partee, and Christina Mulcahy, "Enhancing the Impact of Professional Development in the Context of Preschool Expansion," *AERA Open* 3, no. 4 (2017): https://doi.org/10.1177/2332858417733686.

8. Raj Chetty et al., "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics* 126 (2011): 1593–1660 https://doi.org/10.1093/qje/qjr041; Kenneth A. Dodge, "Community Intervention and Public Policy in the Prevention of Antisocial Behavior," *Journal of Child Psychology and Psychiatry* 50 (2009): 194–200, https://doi.org/10.1111/j.1469-7610.2008.01985.x; Alan B. Krueger, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114 (1999): 497–532, https://doi.org/10.1162/003355399556052.

9. Annalisa Mastri et al., *Evaluating National Ongoing Programs: Implementing the WIA Adult and Dislocated Worker Programs Gold Standard Evaluation* (Washington, DC: Mathematica Policy Research, 2015).

10. Hamre, Partee, and Mulcahy, "Enhancing the Impact."

11. Tseng, Easton, and Supplee, "Research-Practice Partnerships."

12. Douglas H. Clements and Julie Sarama, "Experimental Evaluation of the Effects of a Research-Based Preschool Mathematics Curriculum," *American Educational Research Journal* 45 (2008): 443–94, https://doi.org/10.3102/0002831207312908; Julie Sarama et al., "Scaling Up the Implementation of a Pre-Kindergarten Mathematics Curriculum: Teaching for Understanding with Trajectories and Technologies," *Journal of Research on Educational Effectiveness* 1 (2008): 89–119, https://doi.org/10.1080/19345740801941332.

13. Laurie Miller Brotman et al., "Promoting Effective Parenting Practices and Preventing Child Behavior Problems in School among Ethnically Diverse Families from Underserved, Urban Communities," *Child Development* 82 (2011): 258–76, https://doi.org/10.1111/j.1467-8624.2010.01554.x.

14. Howard S. Bloom and Rebecca Unterman, *Sustained Progress: New Findings about the Effectiveness and Operation of Small Public High Schools of Choice in New York City* (New York: MDRC, 2013); Howard S. Bloom and Rebecca Unterman, "Can Small High Schools of Choice Improve Educational Prospects for Disadvantaged Students?," *Journal of Policy Analysis & Management* 33 (2014): 290–319, https://doi.org/10.1002/pam.21748.

15. Bloom and Unterman, *Sustained Progress*.

16. Bloom and Unterman, *Sustained Progress*.

17. Karl E. Weick and Robert E. Quinn, "Organizational Change and Development," *Annual Review of Psychology* 50 (1999): 361–86, https://doi.org/10.1146/annurev.psych.50.1.361.

18. Weick and Quinn, "Organizational Change and Development."