AERA Open
October-December 2020, Vol. 6, No. 4, pp. 1–21
DOI: 10.1177/2332858420975399
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2020. http://journals.sagepub.com/home/ero

Finding Rigor Within a Large-Scale Expansion of Preschool to Test Impacts of a Professional Development Program

Natalia M. Rojas
Pamela Morris
New York University
Amudha Balaraman
New York City Department of Education

Achieving high-quality preschool at scale is challenging; to do so likely entails a combination of program standards, teacher qualifications and compensation, on-site quality monitoring, and professional development (PD). This study aims to examine the impact of investments in PD within the context of an expansion of universal preschool in one of the nation's largest school districts. We leverage the opportunity provided by a "natural experiment" to estimate PD's effects that embeds an evidence-based math curriculum in interdisciplinary units of study with coaching support on teacher math practices. A total of 95 schools participated in this study (51 treatment and 44 comparison schools). Treatment sites implemented more teacher-led math activities for a longer period compared to control sites. The size and magnitude of the impacts of a curriculum and PD program implemented at scale were comparable to results from studies of small-scale efficacy trials.

Keywords: *preschool, professional development, math*

With growing support among policymakers, universal, publicly funded preschool has expanded over the past decade. In 2016, nationwide, state-funded preschool enrollment reached an all-time high, serving nearly 1.5 million children (Barnett et al., 2017). Proponents of preschool highlight benefits to children's school readiness and later life outcomes, citing research that preschool produces impressive long-term impacts on educational attainment, criminal activity, and health, decades following participation (Belfield et al., 2006; Campbell et al., 2012).

Among all the factors influencing a preschool success, nothing is more important than program quality, namely, teacher-child interactions (Yoshikawa et al., 2013). However, achieving quality at scale is challenging: Evidence indicates that preschool quality may be lower when implemented across a large district, city, or state, especially when compared to those programs implemented under the developers' close supervision or in the context of small, efficacy trials (Dodge, 2009; Dusenbury et al., 2003). High-quality preschool implementation is possible, with a few promising examples in contexts as diverse as Boston, Massachusetts, and Tulsa, Oklahoma (Gormley et al., 2005; Weiland & Yoshikawa, 2013). Results from these studies indicate that assuring high-quality programs at scale entails (1) a professional development (PD) system to support the workforce and (2) the use of an evidence-based curriculum (Gulamhussein, 2013; Weiland et al., 2018). Nonetheless, the effectiveness of this two-pronged approach to improving teacher practices may depend on the quality of programs before implementation.

This study aims to examine improvements to program quality due to PD investments within the context of a universal preschool program serving an ethnically-, racially-, and linguistically diverse sample of children across a range of auspices (e.g., schools, centers) in the nation's largest school district. We leverage a "natural experiment" to estimate the first-year effects of a PD program that embeds an evidence-based math curriculum with coaching support. To understand whether differences in program quality may influence PD impacts on teacher practices, we test for the moderation of program quality before implementation. This study's context—implementation by a major urban school district—represents the real-world conditions of an effectiveness trial and provides evidence regarding whether PD can produce measurable impacts on teacher outcomes amid the complexity of a large public school system. This work is vital in the current early learning policy context where researchers are examining the best way to scale up early learning interventions and the conditions under which preschool is most effective (e.g., Bloom & Weiland, 2015; Duncan & Vandell, 2012).

Improving Preschool Quality Through PD

For preschool programs to eliminate educational opportunity gaps, the programs must be *high-quality* (Barnett, 2011). Among all the factors that influence preschool quality, providing PD to improve teachers' knowledge and practice is particularly important (Hamre et al., 2017). However, the types of PD that most preschool teachers likely receive are not focused on evidence-based teaching practices, are insufficient in terms of duration and intensity, and are not presented in a format that will support sustained changes in teaching practices. Nonetheless, several PD models demonstrated impacts on teaching practice, not only as part of research-led studies but in practice-led, scaled-up implementations as well (e.g., Early et al., 2017). Some examples of PD models tested in smaller efficacy trials and then at scale include Making the Most of Classroom Interactions, My Teaching Partner, and Opening the World of Learning (e.g., Early et al., 2017; Pianta, Mashburn, et al., 2008; Weiland & Yoshikawa, 2013). Implementation and impact results from these studies and others helped define the critical components of a PD program needed to ensure PD is effective.

Specifically, a combination of training and coaching and the use of developmentally appropriate curricula are the key components that produce the largest improvements in teacher's practices, classroom quality, and a range of child outcomes when expanding a PD program to scale (Sarama et al., 2012; Weiland & Yoshikawa, 2013; Weiland et al., 2018). To improve teaching practices and to support gains in children's learning, a PD program should target specific teaching practices (Desimone & Garet, 2015; Zaslow et al., 2010) using an evidence-based curriculum. Furthermore, PD should include a didactic instruction (e.g., workshops) with weekly/biweekly support from coaches (e.g., Bierman et al., 2008; Clements & Sarama, 2011; Morris et al., 2013). The literature suggests that fidelity to the PD program matters, but it is hard to achieve (Durlak & DuPre, 2008). Three dimensions of PD program fidelity are (1) dosage (an index of the quantity of delivery), (2) quality (a measure of the skill with which teachers deliver the material and interact with children), and content (the extent to which the PD was delivered as prescribed).

We examine a PD program, Explore, which combines curricular materials and supports for teachers and leaders. The curricular materials include an evidence-based math curriculum called Building Blocks (BB; Clements & Sarama, 2008; Sarama et al., 2008) and a research-based Pre-K for All (PKA) Interdisciplinary Units of Study (Units[1]) developed by the Division of Early Childhood Education at New York's Department of Education (DECE-DOE). The decision to create a PD track that incorporated a math-focused curriculum was based on evidence that (1) preschool children's math skills are foundational for a broader set of outcomes, including language, reading, and executive function (e.g., Duncan et al., 2007; Watts et al., 2014); (2) preschool math instruction is likely minimal—in terms of the dosage of math instruction and the quality of instruction (Ginsburg et al., 2008); and (3) preschool children's math competencies can increase by training teachers using an evidence-based curriculum with supports (Clements & Sarama, 2007; Presser et al., 2012). Furthermore, because a researcher-led efficacy study of BB called Making Pre-K Count (MPC) in New York City (NYC) took place a few years earlier, there was increased interest in expanding to more sites. Several teachers, leaders, and coaches who participated in the researcher-led efficacy trial played a leadership role in developing and implementing the Explore track. Despite constraints of taking the PD program to scale, many implementation decisions for Explore were based on lessons learned from MPC (e.g., wait until the second year of implementation to focus on math learning trajectories due to the content's complexity) and overseen and delivered by coaches who participated in the efficacy trial. As such, this study serves as an example of how to move a PD program from an efficacy trial to implementation at scale.

The Moderating Role of the Classroom Quality

Although research suggests that a comprehensive PD program, including workshops, coaching, and a curriculum, can change teacher practice, an important question is whether the PD program's effectiveness depends on teachers' skills before implementation, namely, the classroom quality. There is a set of general domains of classroom quality (or teacher-child interactions) that reflect responsive teaching, including emotional support, classroom organization, and instructional support (Hamre, 2014). Irrespective of the content of instruction, high-quality teachers use these general domains of support to engage with children, recognize their needs, and respond in individualized ways that foster social, behavioral, and academic development (Pianta, La Paro, & Hamre, 2008). Often, the introduction of a curriculum, particularly a content-specific one, assumes teachers are competent in delivering high-quality teacher-child interactions, and instead, the curriculum provides support for teachers to target children's specific academic or behavioral skills through instruction in small or large groups, or individually (Wasik & Hindman, 2011). Unclear is whether a PD program's impacts, including a curriculum, vary across initial teacher practice quality (Hamre et al., 2014).

Plausibly, a high-quality classroom before implementing a content-specific curriculum may enhance the quality of the PD program's outcome—in the case of this study, teacher's math practices. That is, teachers who had higher quality classrooms before implementing a math-specific PD program may better incorporate the new curriculum into their existing practice, resulting in higher quality math practices at the end of the year (Goble & Pianta, 2017). Conversely, teachers with initial low levels of quality who struggle with providing classroom organization, emotional, or instructional support to children may find it challenging to implement a new content-specific curriculum, resulting in lower quality math practices. For instance, a teacher who struggles to keep children on task during content instruction may not implement the math practices taught at a PD workshop (e.g., Koth et al., 2008). Furthermore, if a teacher lacks general classroom quality skills, then a coach may spend time helping teachers lay the groundwork (e.g., setting up the classroom) rather than in implementing the new practices (e.g., Morris et al., 2013). Thus, in the context of a PD program at scale and, potentially, limited resources, it is of considerable interest to determine whether the receipt of a PD program that includes a content-specific curriculum is sufficient across initial levels of classroom quality to understand how to support teachers' practices and who might benefit the most from the PD program.

Challenge of Developing and Evaluating a High-Quality PD System at Scale

Nonetheless, carefully designed, well-funded PD programs may fail to affect teacher practice when implemented at a large scale (e.g., Markussen-Brown et al., 2017; Piasta et al., 2017). For instance, the results from the experimental study of classroom size reductions, Tennessee Star, led to widespread implementation of policies and studies of similar interventions; however, the results from subsequent studies were mixed due to various contextual factors, like lack of qualified teachers (Whitehurst & Chingos, 2011). Similarly, studies of school size reductions sparked by notable effects from several small studies did not produce significant changes in achievement results as hoped (Leithwood & Jantzi, 2009). Both of these examples suggest that educational reforms may not have expected impacts when implemented at scale due to factors like poor implementation or differences in contexts.

A second challenge is *how* to rigorously assess an education intervention's impact at scale (Murnane & Willett, 2010). The gold standard design for estimating impacts involves random assignment to a treatment or control group, whereby any differences in outcomes between the two groups can be attributed to the intervention. Randomized control trials in school districts can be challenging to carry out due to the difficulty of obtaining the consent of participants and educational institutions; Thus, randomized designs in large-scale contexts are limited. In recognition of this challenge, researchers investigating preschool impacts use a variety of nonexperimental methods to estimate causal effects, including propensity score matching (e.g., Magnuson et al., 2007) and regression discontinuity (e.g., Gormley et al., 2005; Hustedt et al., 2007; Weiland & Yoshikawa, 2013). Though these nonexperimental designs are not without potential biases (Lipsey et al., 2015), they have the advantage of being recognized as relatively strong designs being easily applied to programs at scale.

Another compelling strategy, in the absence of randomization and regression discontinuity design opportunities, is to rely on natural experiments, namely, a study where researchers take advantage of a situation in which two otherwise identical groups are affected differently (i.e., exposed to a treatment and control condition) by a "natural" event that is exogenous to the treatment and the outcome (Murnane & Willett, 2010). The process governing the exposure to the different conditions resembles random assignment and creates otherwise identical pretreatment groups (Lipsey et al., 2015). Numerous research examples use such "natural" variation in policies or events to produce unbiased estimates of the effects. For example, Hoxby (2001) estimated the effects of school vouchers on school choice using school district boundaries determined by streams. Similar to in this study, researchers took advantage of naturally occurring pockets of randomization within school lotteries that mimic random assignment. For instance, if more students apply than there are seats available, within priority groups established by schools or districts, a lottery is used to choose which students are offered a seat randomly (e.g., Dobbie & Fryer, 2011; Lipsey et al., 2018; Unterman et al., 2016). The use of nonexperimental approaches, including natural experiments, constitutes a research design for addressing casual questions while meeting the methodological and ethical bars that are implicit in research studies at scale.

Present Study

The translation from efficacy studies to implementation across large preschool systems often requires adaptations from nonresearchers based on local constraints (e.g., number of PD days; funds) that may not match what was done in research studies. Although there are other studies of the BB curriculum, using more rigorous designs and measuring child outcomes, this study is one of the few studies that improve our understanding of PD impacts on teachers' math practices as a function of participation in real-world, large-scale PD effort. Studies like this one are essential in understanding the success, or lack thereof, of educational approaches when adopted and implemented at large scale. We leverage a natural experiment within the NYC system of assigning sites to PD programs that resulted from a delay in funding decisions outside of the DECE-DOE, preschool programs, and researchers' control. We address the following questions:

Research Question 1: What is the impact of the PD track, Explore, on the amount and quality of math instruction in preschool classrooms after 1 year of implementation?

Research Question 2: Do baseline classroom quality scores moderate the intervention's effects on the amount and quality of math instruction?

Method

Launched in 2014, PKA represents NYC's commitment to providing free, full-day, high-quality preschool to every 4-year-old (about 68,500 preschool seats in 1,850 sites). As part of their commitment to quality, DECE-DOE began building a system of PD to include the central features linked to quality—training, coaching, and curricula.

Track Assignment Procedures

A set of procedures were put in place that led to the assignment of sites to PD tracks outside of the control of the DECE-DOE, programs, and researchers; this study leverages a natural experiment in the assignment process of sites to PD tracks. In the spring of 2016, there were four PD tracks: (1) Explore, an evidence-based math curriculum, and the Units developed by the DECE-DOE that supports high-quality teacher practice and children's development across domains by integrating math concepts into the classroom; (2) Create, an arts-based approach to incorporate visual arts, dance, theater, and music into ongoing instruction to promote learning across domains; (3) Thrive, an evidence-based set of strategies for supporting children's social-emotional development and behavioral regulation, as well as for supporting family engagement; and (4) Inspire, a series of topics aligned with the district's quality standards that support DECE-DOE instructional and child development goals. The Explore PD track was made possible by a mix of funding sources (public and private/external), which had not yet been finalized at the time of the PD track assignment.

Assignment to PD tracks was based on the following two conditions: (1) site leaders' rank-order preference for Explore, Create, and Thrive,[2] and (2) recommendations for the PD tracks from a social worker and/or instructional coach working with each site. An algorithm was developed to assign sites to PD tracks. First, the algorithm created six priority groups based on site preference and recommendations for each PD track (see Appendix A for a description of priority groups). A site could be in a different priority group for each track based on their preference and recommendations ranking (e.g., a site could be in priority Group 1 for Explore and priority Group 5 for Create). Second, sites were assigned to a PD track in the order of their priority group until the PD track capacity was met, beginning with the Thrive, Create, and Explore tracks. Each PD track (except Inspire) had a maximum capacity of sites that could be served. When demand for a PD track exceeded the capacity of that PD track (i.e., oversubscribed), the site assignment algorithm randomly assigned sites from within a priority group to either that PD track or the Inspire PD track (essentially a lottery; e.g., if the number of sites in Create's priority Group 4 exceeded the maximum number of sites allowed in Create, then sites in priority Group 4 were randomly assigned to the Create or Inspire). Sites not assigned to Thrive, Create, or Explore due to oversubscription were also placed in the Inspire track.

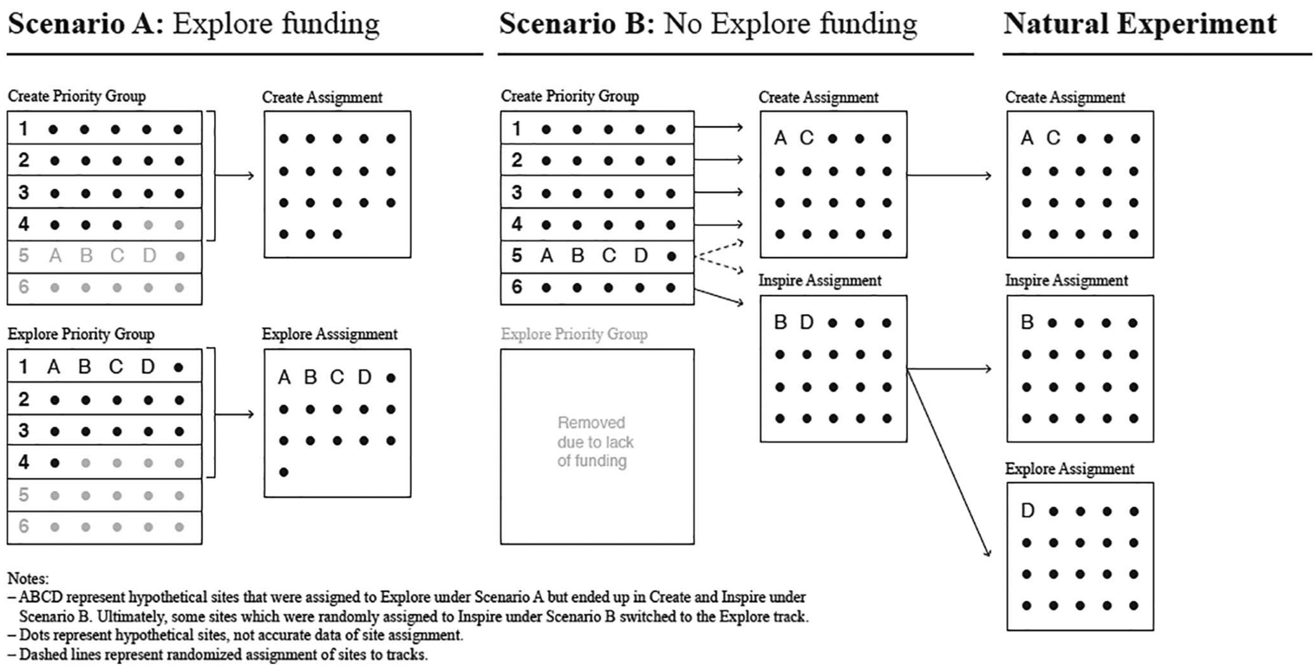
How Sites Were Assigned to Tracks in the Spring of 2016

In Spring 2016, with funding for the Explore PD track uncertain, the assignment algorithm was run twice to create two site assignment lists. The first site assignment list (Scenario A) is based on the condition that Explore funding was available and sites could be assigned to Explore, Create, Thrive, and Inspire. The second site assignment list (Scenario B) is based on the condition that Explore funding was not secured and, thus, sites could be assigned only to Create, Thrive, and Inspire (see Figures 1 and 2). Since the site assignment algorithm for Scenario B used the same priority groups from Scenario A, this meant that sites that would have been assigned to Explore in Scenario A tended to be in a lower priority group for Create. Because demand for the Create track, in Scenario B, exceeded the number of spots available, a set of sites were randomly assigned within a priority group to the Create or the Inspire PD track in Scenario B.

After the site assignment algorithm created the two lists of site assignments (Scenarios A and B) and because funding for the Explore PD track remained uncertain, the DECE-DOE notified sites of their PD track assignment based on the results from Scenario B. However, several weeks later, funding for an additional cohort of Explore was secured. This meant that the sites that would have received Explore in Scenario A needed to now be reassigned from their current track (i.e., the Scenario B track assignment they had already received—Create or Inspire) to the Explore track. Sites already assigned to Create but that would have been assigned to Explore under Scenario A did not change PD track assignment to Explore because of the need for a certain number of sites in the Create track. As a result, only sites assigned to Inspire but that would have been offered Explore under Scenario A were eligible to participate in Explore.

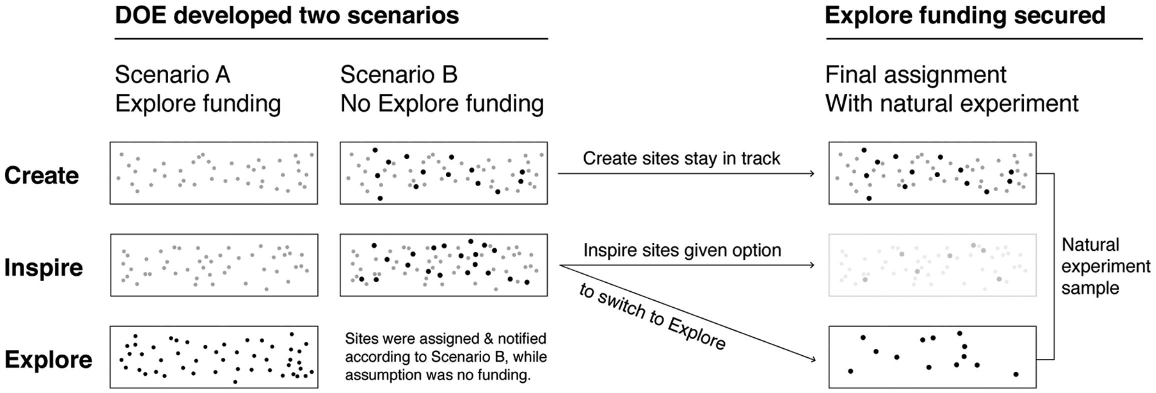
Thus, during the 2016–2017 school year, two groups of sites in the PKA system were eligible for Explore and would have been assigned to Explore, but only *one* received the offer to participate in Explore (due to late funding). The treatment group included sites that would have been assigned to Explore in Scenario A and were ultimately assigned to receive Explore after all. Our comparison group comprised sites that would have been assigned to Explore in Scenario A but were ultimately placed in Create. This comparison group allowed us to examine what might have happened to our Explore (treatment) group, on average, if they had been assigned to the other tracks.

Examining potential ways that the site assignment process may result in the treatment and control groups that were different from one another in observed and unobserved ways is necessary. As emphasized by Cook and Campbell (1979) and Dunning (2008), assignment to treatment and control



Notes:
– ABCD represent hypothetical sites that were assigned to Explore under Scenario A but ended up in Create and Inspire under Scenario B. Ultimately, some sites which were randomly assigned to Inspire under Scenario B switched to the Explore track.
– Dots represent hypothetical sites, not accurate data of site assignment.
– Dashed lines represent randomized assignment of sites to tracks.

FIGURE 1. *Hypothetical track assignment process within priority groups resulting in the natural experiment.*



Sites were assigned & notified according to Scenario B, while assumption was no funding.

FIGURE 2. *Overview of the process of the natural experiment.*

conditions—here, the Explore PD track—must be *as if* random. This implies that the Explore track assignment is independent of observable and unobservable factors that might influence teachers' math practices. Furthermore, the treatment and control groups must be balanced concerning measurable variables that might explain teachers' math practices. Particularly important, sites should not appear to self-select into their PD track in ways that might be associated with a propensity to teach math. Perhaps most concerning is that our treatment group is made up of sites that would have received Explore in scenario A but were given the *option* to be reassigned to Explore; this means that our treatment group is made up of sites that *chose* to be reassigned to Explore. In contrast, sites that would have gotten Explore in Scenario A but ended up in the Create track (ultimately, our control group) did not have the option to be reassigned because of funding constraints.

Nonetheless, we argue that the site preference data, which predate PD track assignment, reflect their true interest in a track, and as such, sites that ranked Explore as a top choice would have chosen to be reassigned to Explore. Furthermore, in focus groups, site leaders described making decisions about ranking site preference without consulting teachers. Taking that into consideration, and the fact that social workers/instructional coordinators recommendation played a crucial role in the PD track assignment process, suggests that teachers, who are the actual recipients of the treatment and control conditions, were divided into the two PD tracks most often without their choice or knowledge; thus, teachers were unlikely to self-select into either group in a way that might influence their propensity to implement math. A series of robustness checks are conducted to test these assumptions (Appendix B).

Program Services

Two years of program services were provided. The Explore program consisted of curricula, including BB (Clements & Sarama, 2008) and the Units, training, and on-site coaching. The BB curriculum is a multifaceted learning activity sequence targeting numeric and geometric or spatial topics laid out across 30 weeks in an easy-to-read, scripted manual. Curricular activities are organized based on the natural progressions by which children learn and develop math competencies over time (Clements et al., 2013). The following support was provided: (1) professional learning delivered by the developers of BB and expert facilitators trained in the BB curriculum, teachers participated in 4 full-day trainings (6 hours per day), and leaders participated in 3 half-day (4 hours per day) training across the school year, and (2) on-site coaching for teachers by external coaches trained by BB certified trainers. On average, coaches observed the teacher in the classroom once a month for 1 hour and employed strategies such as modeling, providing feedback, and discussing implementation with the teachers. Coaches debriefed with leaders during their visits, and they served as the direct point of contact as for questions about Explore implementation. The amount of PD offered to teachers and leaders in this study was less than other BB studies. In a study in NYC of BB, teachers participated in 6 days of training in the first year of implementation and 3 hours of in-classroom coaching each week (Morris et al., 2016).

Teachers in the non-Explore track attended 4 full-day trainings and leaders attended 4 half-day trainings across the school year that focused on incorporating visual arts, dance, theater, and music to promote children's learning. Instructional coordinators and/or social workers supported teachers in the classrooms. The dosage of coaching by instructional coordinators and/or social workers was dependent on need (e.g., low site quality scores), and changed across the year.

Sample

Ninety-five schools participated (51 schools in Explore and 44 in non-Explore; see Table 1) in the 2016–2017 school year (the first year of implementation). The 95 PKA programs included 32 public district schools, seven Administration for Children's Services–NYC Early Education Centers (ACS-NYCEECs), 52 DOE-NYCEECs, and four PreK Centers. District and preK center teachers must have a New York State teaching certification in early childhood along with a bachelor's degree. NYCEEC teachers must have a teaching license or certificate in ECE/ECE students with disabilities or a bachelor's degree with a plan for obtaining Early Childhood Education certification. NYCEEC teachers must commit to earning a New York State teaching certification within 3 years of their start date as a pre-K lead teacher.

The children who attended participating sites were racially, linguistically, and socioeconomically diverse (Table 1). Across the sites, 34% of the children were Hispanic, 28% were Black, 15% were White, 20% were Asian, and 3% were of mixed, or other, race. Fifty-one percent of children were female. Thirty-three percent spoke a language other than English at home. Approximately 53% of children were eligible for free or reduced-price lunch, and 7% had an individualized education plan (IEP).[3]

Classroom Observation Protocol

Ten trained graduate student observers (blinded to sites' intervention status) collected classroom-level data. Observers underwent training with the classroom observational tool developer, including a practice observation in a site not enrolled in the study. The observational tool was collected in one randomly selected classroom per site at one time point in the spring (March–May). Classroom observations were scheduled on days when coaches were not present. Observations were completed "live" on-site over the first 3 hours of the day. When observations were double-coded for reliability purposes, the resultant data utilize the mean scores averaged between coders.

Measures

Classroom Outcomes. The Classroom Observation of Early Mathematics–Environment and Teaching (COEMET), a classroom observation tool that measures math instruction, was developed based on the characteristics and teaching strategies of effective teachers of early childhood mathematics (e.g., Clements & members of the Conference Working Group, 2004). The COEMET has two main sections, Classroom Culture and Specific Math Activities (SMA). Assessors complete the Classroom Culture section once to reflect their entire observation. The Classroom Culture section, which assesses teachers' general approach to math education, includes items on the mathematical environment and interactions (e.g., environment showed signs of math, children's math work on display, teacher actively interacted) and on the personal attributes of the teacher (e.g., teacher was knowledgeable about math, showed math learning could be enjoyable, showed curiosity for math). They complete an SMA form for each teacher-led formal math activity, defined as an activity led by a teacher that lasted at least 30 seconds; developed math knowledge; had a discernible topic, goal, and task; and involved multiple conversational turns between a teacher and a child. Observers completed a mini SMA form to document when a "simple" or "routine" math activity (e.g., singing a song about numbers) led by a teacher that does not include an extensive conversation about math content occurred. Interrater reliability for the COEMET,

TABLE 1

Baseline Equivalence of the Natural Experiment Sample

Characteristics	Explore, %	Non-Explore, %
Site type		
Public schools	41	$25^{\dagger}$
DOE-NYCEEC	47	64
ACS-NYCEEC	4	11
PreK center	8	$0^{\dagger}$
Child demographics		
Race		
Asian	18	22
Black	36	20
Hispanic	32	36
White	11	$19^{\dagger}$
Multirace	3	3
Female	51	51
Poverty	55	51
With individualized education plan	8	6
Language other than English spoken at home	30	35
Program quality		
Early Childhood Environment Rating Scale–Revised	3.93	4.28*
CLASS Emotional Support	6.27	6.25
CLASS Classroom Organization	6.11	6.08
CLASS Instructional Support	3.15	3.39
Joint test of all baseline characteristics	$F(19, 73) = 1.43, p = .14$	
Sample size		
Sites	51	44

Note. DOE-NYCEEC = Department of Education–New York City Early Education Center; ACS-NYCEEC = Administration for Children's Services–New York City Early Education Center; CLASS = Classroom Assessment Scoring System.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

computed via simultaneous classroom visits by pairs of observers (10% of all observations, with pair memberships rotated), was 89%; 96% of the disagreements were the same polarity (i.e., if one agreed, the other was strongly agree).

We created two outcome variables to represent the number of math activities observed in each classroom: (1) the count of teacher-led math activities represents the total number of teacher-led math activities observed (i.e., activities recorded on the SMA and mini SMA that were led by teachers) and (2) the minutes of teacher-math activities represent the total number of minutes of teacher-led math activities observed.

We created three variables that represent the math quality observed: (1) classrooms with at least one observed teacher-led math activity, (2) average math quality scores, and (3) dichotomous moderate math quality score. Ratings on the quality of math instruction are available only in classrooms where math was observed. First, we created a dichotomous variable that indicates whether a classroom had at least one observed teacher-led math activity (0 = *no teacher-led math activities were observed*; 1 = *at least one teacher-led math activity was observed*). For classrooms where a teacher-led math activity was observed, the average math quality score was calculated by averaging across the items and then averaging across math activities for the final score to create the average math quality score. However, since the number of classrooms where at least one teacher-led math activity was observed was different between program and control groups (71% vs. 34%), this variable does not represent the true impact; as such, we created a variable that accounts for the fact that some classrooms are missing a math quality score (because they were not observed implementing a teacher-led math activity). This variable, dichotomous moderate math quality score, was calculated for all classrooms, regardless of whether they were observed conducting a math activity. Classrooms were considered to have moderate-quality math instruction, and thus were coded 1, if they had an average math quality score at or above a rating of 2 (median split) on a scale from 1 to 5. Classrooms were coded 0, or low-quality math instruction, if they had an average math quality rating below 2 or no quality rating. Finally, the Mathematical Classroom Culture score was computed by summing the items from the Classroom Culture section.

Treatment Variable. A dummy variable was created to represent assignment to the treatment or comparison condition (Explore = 1; comparison = 0).

Covariates, Moderators, and Descriptive Characteristics. The covariates were entered at the school level (i.e., at the level of assignment to track), as is recommended in the analysis of cluster-randomized trials and as the data were available from administrative records (Raudenbush et al., 2007).

Classroom quality. To control for the variability in sites' classroom quality, baseline measures of the Early Childhood Environment Rating Scale–Revised (ECERS-R; Harms et al., 2003) was included as a covariate. The Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008; collected within the past 3 years) was included as a covariate and a moderator. CLASS observations were conducted once every 3 years. The NYC DOE used a modified version of the CLASS protocol where the number of classrooms and the number of cycles each classroom was observed varied depending on the site's size. All CLASS scores were analyzed and reported at the site level. Each of the three domains of the CLASS were separately examined as moderators in analyses for Research Question 2. The Emotional Support domain reflects the extent to which teachers support the classroom's emotional and social functioning. The Classroom Organization domain reflects processes related to children's behavior, time, and attention. The Instructional Support domain refers to how teachers encourage higher order thinking and facilitate children's use of language. In the current study, Cronbach's alphas were the following: Emotional Support (.84), Classroom Organization (.87), and Instructional Support (.93).

Administrative data. From records, we obtained information on child demographics (site proportion), gender (school proportion), percentage of children at the site who received free/reduced-price lunch, percentage of children at the site who had an IEP, and percentage of children who come from homes that spoke a language other than English. We controlled for site characteristics that are considered relevant in the NYC context: borough and site type (district school, DOE-NYCEECs, ACS-NYCEECs, PreKCenter). We used a vector of dichotomous indicators to represent borough and site type, each coded 1 when the site was located in a particular borough or located within a site type, 0 otherwise. These covariates predict children's early cognitive and educational outcomes in other studies, and there is a consensus in the preschool literature that these should be controlled in impact analyses (Clements et al., 2011; Wong et al., 2008).

Analytic Plan

Our design relies on a treatment-on-treated design, effectively comparing sites that were offered and agreed to take up the offer of a change of assignment of PD track, relative to those who were not offered to change their PD track assignment. As such, our treatment-on-treated design estimates the impact of receiving Explore PD on the "treated" sites. We approach our first research question, about the impact of Explore on the amount and quality of math instruction, with a series of ordinary least squares regression models, with standard error correction (Huber-White) for clustering of classrooms within sites (an approach more commonly used by economists in cluster-randomized trials; Murnane & Willett, 2010). Consistent with consensus in the field, we interpret effects in the 0.10 to 0.30 range as small, effects in the 0.30 to 0.60 range as moderate, and effects in the 0.60 and higher range as large (Hill et al., 2008). For all analyses, we used the STATA statistical software package. For Model 1:

$$\begin{aligned} Y_i = {} & \alpha + \beta_1\left(DEMOGRAPHICS\right) + \beta_2\left(SITETYPE\right) + \\ & \beta_3\left(BOROUGH\right) + \beta_4\left(PRE\text{-}TEST\ QUALITY\right) + \\ & \beta_5(RX) + \varepsilon. \end{aligned}$$

Y*i* represents the outcome variables of interest at the classroom level, *i* represents *classrooms; DEMOGRAPHICS* is a vector of child demographics at the school-level (percentage of children receiving free/reduced-price lunch, percentage of children with IEPs, percentage of children who speak a language other than English at home, percentage of each of these races (Black, Hispanic, Asian, White, and multirace) enrolled in each school, and gender); *SITETYPE* is a set of dummy variables that represent the type of school setting (ACS-NYCEEC, DOE-NYCEEC, District School, or Prek Center); *BOROUGH* is a vector of five dummy variables for each of the boroughs that the programs were located in; *PRE-TEST QUALITY* is a vector of beginning pretest quality scores of CLASS and ECERS; *RX* is the intervention status. Baseline variables are limited to the few static demographic and school characteristics available in the DECE database. Differences between the baseline characteristics of Explore and non-Explore group are examined, to determine whether randomization "worked."

To address the aims of Research Question 2, we separately reestimated our models (described earlier) and added an interaction term, calculated as the product of the baseline score on each of the CLASS subscales and intervention status (Explore = 1, non-Explore = 0).

Results

We explored the extent to which the natural experiment sample yielded comparable treatment and control groups. We conducted *t* tests by treatment at the site level on the following child- and site-level pretest characteristics: children's race, gender, site type, percentage of children with IEPs, language other than English spoken at home, free/reduced-price lunch, and baseline program quality. These analyses yielded only one statistically significant difference at the .05 level across 17 variables tested. Comparison sites were statistically more likely to have higher ECERS scores compared to the intervention sites; however, notably, this statistically significant difference should bias estimates of treatment impact downward (given they suggest that control sites were of slightly higher quality). All other tested baseline characteristics were similar across groups (Table 1). This pattern suggests that the assignment to conditions was successful in producing comparable groups for assessing treatment impact. Tables 2 and 3 show the correlations across study variables.

What Is the Impact of Explore on the Amount and Quality of Math Instruction?

Table 4 summarizes the impacts on teachers' math practices at the end of the first year of implementing Explore. We found positive impacts on five out of the six outcome variables. Intervention effects were positive and statistically significant for count of teacher-led math activities (effect size [*ES*] = 2.73, $p < .001$), minutes of teacher-led math activities (*ES* = 1.78, $p < .001$), % of classrooms with at least one observed teacher-led math activity (*ES* = 1.58, $p < .001$), dichotomous moderate math quality score (*ES* = 0.95, $p < .001$), and the mathematical classroom culture scale (*ES* = 2.15, $p < .001$). The effect sizes were relatively large (with *ES*s = 0.95–2.73 across measures). In Explore and non-Explore sites, most of the teacher-led math activities were focused on number concepts. The teaching of operations, geometry, patterning, and measurement were at lower levels in both the Explore and non-Explore sites—on average, less than one activity focused on each of these math areas. Explore teachers were observed to deliver statistically significantly more activities within each of the following math content areas compared to non-Explore sites ($p < .05$): number, operations, patterning, and measurement (see Table 5). The *ES*s ranged from 0.64 to 1.38.

Does Baseline Classroom Quality Score Moderate the Effects of the Intervention on the Math Practices?

Only two interactions were statistically significant (see Table 6). An interaction was detected between baseline classroom organization and intervention status on the count of math activities ($b = 0.28$, $p = .00$). This suggested that sites with lower baseline classroom organizational skills conducted more teacher-led math activities at the end of the first year of implementation if they were in Explore and sites with higher baseline classroom organization skills conducted more teacher-led math activities at the end of the first year of implementation if they were in Explore (see Figure 3). Inspection of the simple slopes revealed that sites with high classroom organization (1 *SD* above the mean), the average number of teacher-led math activities for those assigned to Explore was 3.44 *SD* higher than those in the non-Explore condition. For sites with low classroom organization (1 *SD* below the mean), the average number of teacher-led math activities for those assigned to Explore was 2.93 *SD* higher than those in the non-Explore condition.

A second interaction was detected between baseline emotional support and intervention status on the mathematical classroom culture score ($b = 0.83$, $p = .00$). This suggested that baseline emotional support moderated the relation between intervention status and mathematical classroom culture. Figure 4 illustrates that receiving the Explore PD was significantly related to higher mathematical classroom culture among sites with higher baseline emotional support (0.46 *SD* higher than those in the non-Explore condition). In contrast, for lower emotional support sites, receiving the Explore intervention did not significantly predict the mathematical classroom culture scores (simple slope = ns).

Discussion

This study investigated the impacts of an at-scale PD program on preschool teacher's practices. We examined such impacts within the context of a natural experiment, with PD implemented under real-world conditions (Institute of Education Sciences & National Science Foundation, 2013). The advantage of this approach is that we determined whether an at-scale, district-sponsored PD, as authentically implemented, resulted in the intended outcomes. The results are essential to consider, within the context of PD at scale, given that the PD model was developed following research-based recommendations for effective PD and the similar investments (e.g., time, financial) currently being made in PD programs across the country.

We found impacts on the number of minutes (13.26 more minutes in Explore) and the count of math activities (1.55 more math activities in Explore), which were substantially larger than those seen in previous studies of BB, where the program group typically spent 2 to 5 more minutes on math instruction than the control (Clements & Sarama, 2008; Morris et al., 2016). The size and magnitude of our findings are comparable to the MPC study, which took place in NYC in 2013–2015 and utilized a similar observation procedure. The mean number of activities and minutes of math were less for both the Explore and non-Explore groups (2.53 activities,19.67 minutes in Explore; 0.98 activities, 6.09 minutes in non-Explore) compared to the intervention and control groups in MPC (5.94 activities, 46.80 minutes in the treatment group; 4.37 activities, 34.85 minutes in the control group). Our findings suggest that the PD supports did not

TABLE 2

Correlations Between Study Variables (Total Sample)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1. ECERS (BL)	1.00																									
2. Emotional Support (BL)	.07	1.00																								
3. Classroom Organization (BL)	.02	.8 ***	1.0																							
4. Instructional Support (BL)	.05	.33**	.7*	1.0																						
5. % Asian	.19	.08	.02	−.07	1.00																					
6. % Black	−.22*	.00	.00	−.04	−.49***	1.00																				
7. % Hispanic	−.07	−.08	−.01	−.02	−.33***	−.40***	1.00																			
8. % White	.16	.00	−.01	.13	−.07	−.44***	−.20*	1.00																		
9. % Multiracial	.12	.06	.07	.10	−.08	.01	−.15	.09	1.00																	
1. % Females	.06	.24	.32**	.16	−.17	.20*	.03	−.15	.07	1.00																
11. % FRL	−.26**	.00	.03	−.20	−.32***	.27*	.38***	−.45***	−.15	.02	1.00															
12. % IEPs	.00	−.05	−.06	−.08	−.04	−.06	.02	.11	.03	−.04	−.13	1.00														
13. % LOTE	.07	−.07	−.05	−.17	.56***	−.58***	.28**	−.12	−.18	−.19	.12	.04	1.00													
14. Brooklyn	−.06	−.01	.10	−.03	−.09	.27**	−.33***	.13	−.06	−.15	.02	.07	−.15	1.00												
15. Manhattan	.01	.00	−.01	.01	−.18	.10	.14	−.12	.09	.00	.14	.27**	−.08	−.26**	1.00											
16. Queens	.23*	.14	.07	.20	.39***	−.27**	−.09	.02	.09	.10	−.28**	−.15	.21*	−.57***	−.26**	1.00										
17. Staten Island	.00	.10	−.03	−.16	−.05	−.07	−.05	.19	.17	.08	−.07	.13	−.12	−.11	−.05	−.11	1.00									
18. Bronx	−.24*	−.23	−.20	−.17	−.23*	−.06	.48***	−.18	−.18	.04	.26**	−.18	.03	−.31***	−.14	−.32***	−.06	1.00								
19. Count of math activities	.07	.11	.12	−.09	.01	.05	−.07	−.02	.09	.01	.05	.11	−.01	.24*	−.09	−.21*	.06	.02	1.00							
2. Minutes of math	.06	.12	.14	.10	.14	.09	−.13	−.19	.19	.00	.03	.03	.06	.13	−.09	.00	−.03	−.08	.67***	1.00						
21. % With at least 1 teacher-led math activity	−.09	−.02	−.04	−.10	.03	.15	−.05	−.20	.05	−.04	.09	.07	.01	.17	−.09	−.25**	.14	.15	.64***	.65***	1.00					
22. Dichotomous moderate math quality	−.03	.39*	.29	.31	−.09	.03	.16	−.11	−.20	.16	.02	−.05	−.10	−.18	−.02	.17	−.02	.06	−.01	.17	.	1.00				
23. Average math quality	.04	.17	.13	.08	.11	.03	−.11	−.05	−.01	.08	−.06	−.03	−.04	−.08	−.09	.04	.06	.10	.31***	.36***	.40***	.63***	1.00			
24. Math Classroom Culture	−.13	.08	.05	.08	−.01	.05	−.01	−.06	.05	.01	−.06	.08	.05	.08	−.07	−.11	.02	.09	.60***	.46***	.55***	.45***	.47***	1.00		
25. Emotional Support (FU)	.07	.11	.22	.32**	.24*	−.34***	.03	.19	.00	.01	−.22*	−.06	.21*	.03	−.06	−.02	.04	.04	−.17	−.17	−.11	−.19	−.08	−.14	1.00	
26. Classroom Organization (FU)	.08	.04	.14	.17	.18	−.38***	.23***	.09	−.03	−.13	−.09	−.14	.21*	−.13	−.06	.02	.04	.18	−.11	−.23*	−.06	−.12	−.08	−.11	.83***	1.00
27. Instructional Support (FU)	−.01	.36**	.35**	.11	.14	−.17	−.08	.18	.10	.08	−.09	−.15	.02	.17	−.15	−.20	.08	.14	.11	−.02	−.02	−.07	−.05	.16	.38***	.32***

Note. ECERS = Early Childhood Environment Rating Scale–Revised; BL = baseline; FU = follow-up; FRL = free/reduced-price lunch; LOTE = language other than English spoken at home; IEP = individualized education plan.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

TABLE 3

Correlations Between Study Variables Broken Down by Intervention Group

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Non-Explore Sample																										
1. ECERS (BL)	1.00																									
2. Emotional Support (BL)	−.18	1.00																								
3. Classroom Organization (BL)	−.21	.77***	1.00																							
4. Instructional Support (BL)	−.11	.43**	.37*	1.00																						
5. % Asian	.26	.05	.08	−.03	1.00																					
6. % Black	−.23	.24	.16	.19	−.39**	1.00																				
7. % Hispanic	−.14	−.13	.00	−.17	−.44***	−.25	1.00																			
8. % White	.11	−.12	−.21	.00	−.22	−.40**	−.29*	1.00																		
9. % Multiracial	−.01	−.02	−.03	.22	−.08	.08	−.23	.12	1.00																	
10. % Females	−.01	.36*	.44**	.17	−.28	.45***	.04	−.22	.00	1.00																
11. % FRL	−.35*	.09	.01	.02	−.43***	.45***	.48***	−.53***	−.07	.22	1.00															
12. % IEP	−.12	−.16	−.08	−.01	−.22	−.23	.23	.22	.25	−.11	−.13	1.00														
13. % of LOTE	.28	−.07	−.07	−.20	.54***	−.53***	.26	−.27	−.30*	−.29*	.02	−.06	1.00													
14. Brooklyn	−.09	.00	.02	.03	−.08	.33*	−.29	.05	−.04	.01	−.03	.03	−.27	1.00												
15. Manhattan	−.09	−.03	.06	.19	−.14	.13	.17	−.18	.11	.06	.11	.48***	−.02	−.18	1.00											
16. Queens	.32	.17	.15	.08	.35*	−.32*	−.19	.16	.03	−.08	−.30*	−.25	.26	−.55***	−.30*	1.00										
17. Staten Island	.01	.04	−.25	−.08	−.12	−.11	−.12	.37*	.09	−.02	−.17	.26	−.15	−.09	−.05	−.15	1.00									
18. Bronx	−.26	−.23	−.17	−.29	−.23	−.01	.52***	−.28	−.12	.05	.42***	−.18	.04	−.25	−.14	−.42**	−.07	1.00								
19. Count of Math Activities	.17	−.16	−.21	.02	.09	−.08	.08	−.13	.20	−.13	−.06	−.08	.17	−.08	−.06	−.06	.00	.22	1.00							
20. Minutes of Math	.06	.06	.01	.18	.35*	−.09	−.14	−.17	.17	−.20	−.14	−.05	.29	−.07	.06	.07	−.04	−.03	.68***	1.00						
21. % with at least 1 teacher-led math activity	.06	−.14	−.31	−.13	.10	−.04	.03	−.14	.23	−.12	−.06	−.04	.09	.03	−.06	−.30*	.21	.34*	.72***	.51***	1.00					
22. Dichotomous mod math quality	−.31	.65*	.54	.46	−.26	−.06	.35	.00	−.10	.30	.26	.02	−.03	−.40	.06	−.04	.19	.28	.14	.17	.00	1.00				
23. Average Math Quality	−.01	.15	−.06	.01	.19	−.11	−.07	−.01	−.01	.10	−.11	.00	.01	−.19	.02	−.08	.28	.21	.53	.43***	.41**	.55*	1.00			
24. Math Classroom Culture	−.03	−.22	−.21	.03	.08	−.11	.02	.01	−.02	−.02	−.11	−.03	.10	−.15	−.11	−.09	.14	.33*	.60***	.33*	.46***	.54*	.59***	1.00		
25. Emotional Support (FU)	.19	.32	.40*	.37*	.27	−.19	−.15	.06	.09	.18	−.34*	−.02	.13	−.09	.10	.14	.00	−.16	.06	.14	.08	.37	.16	.02	1.00	
26. Classroom Organization (FU)	.29	.08	.21	.05	.27	−.31	.02	.01	.10	.04	−.25	−.15	.18	−.23	−.10	.18	.00	.11	.25	.12	.27	.21	.25	.13	.82***	1.00
27. Instructional Support (FU)	.12	.38*	.41*	.23	.31*	−.09	−.14	−.10	−.13	.33*	−.07	−.24	.09	.11	−.28	−.07	.00	.18	.22	−.01	.30	.38	.40**	.36*	.34*	.38*

(continued)

TABLE 3. (CONTINUED)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Explore Sample																										
1. ECERS (BL)	1.00																									
2. Emotional Support (BL)	.28	1.00																								
3. Classroom Organization (BL)	.20	.78***	1.00																							
4. Instructional Support (BL)	.00	.24	.20	1.00																						
5. % Asian	.15	.11	−.05	−.11	1.00																					
6. % Black	−.15	−.22	−.15	−.09	−.58**	1.00																				
7. % Hispanic	−.06	.00	−.02	.09	−.21	−.51**	1.00																			
8. % White	.17	.24	.34	.24	.12	−.48***	−.11	1.00																		
9. % Multiracial	.18	.18	.20	−.06	−.08	−.04	−.10	.08	1.00																	
10. % Females	.10	.10	.18	.14	−.06	.05	.02	−.08	.13	1.00																
11. % FRL	−.21	−.09	.04	−.35	−.21	.13	.29*	−.36**	−.21	−.16	1.00															
12. % IEPs	.09	.02	−.04	−.10	.10	−.04	−.09	.09	−.07	.01	−.15	1.00														
13. % LOTE	−.06	−.05	−.01	−.23	.59***	−.62***	.30*	.02	−.11	−.10	.21	.11	1.00													
14. Brooklyn	.02	−.03	.16	−.01	−.08	.18	−.36*	.33*	−.09	−.27*	.03	.05	−.03	1.00												
15. Manhattan	.07	.06	−.11	−.26	−.21	.06	.13	−.05	.07	−.06	.16	.16	−.11	−.33*	1.00											
16. Queens	.13	.12	.00	.28	.42***	−.16	−.03	−.28*	.16	.27	−.25	−.05	.13	−.56***	−.22	1.00										
17. Staten Island	−.01	.16	.19	−.24	.02	−.03	.01	−.05	.23	.19	.03	.05	−.10	−.13	−.05	−.09	1.00									
18. Bronx	−.25	−.24	−.23	−.05	−.23	−.09	.43***	−.06	−.23	.03	.11	−.18	.01	−.36**	−.15	−.25	−.06	1.00								
19. Count of Math Activities	.19	.29	.33	−.04	.04	−.05	−.12	.22	.04	.11	.08	.11	−.03	.29*	−.15	−.19	.11	−.07	1.00							
20. Minutes of Math	.20	.18	.27	.18	.06	.04	−.09	−.11	.20	.18	.12	−.01	−.02	.13	−.23	.11	−.02	−.11	.59***	1.00						
21. % with at least 1 teacher-led math activity	−.04	.12	.25	.07	.03	.14	−.08	−.18	−.09	.05	.19	.05	.01	.15	−.17	−.09	.09	.01	.54***	.66***	1.00					
22. Dichotomous mod math quality	.10	.19	.08	.25	.03	.00	.11	−.16	−.21	.08	−.09	−.11	−.10	−.17	−.05	.26	−.11	.02	−.13	.11	.00	1.00				
23. Average Math Quality	.13	.19	.30	.19	.08	.05	−.12	−.03	−.01	.07	−.05	−.08	−.04	−.08	−.17	.23	−.11	.03	.16	.26	.34**	.67***	1.00			
24. Math Classroom Culture	−.08	.38*	.27	.27	−.04	.01	.02	−.01	.08	.07	−.07	.06	.08	.13	−.08	−.01	−.06	−.08	.53***	.42***	.52***	.38**	.35*	1.00		
25. Emotional Support (FU)	−.03	−.05	.09	.22	.21	−.40***	.15	.30*	−.03	−.14	−.13	−.05	.25	.15	−.16	−.22	.07	.18	−.19	−.29*	−.16	−.28	−.19	−.17	1.00	
26. Classroom Organization (FU)	−.04	.02	.08	.15	.10	−.39***	.37**	.13	−.08	−.27	.03	−.11	.20	−.02	−.02	−.16	.07	.22	−.17	−.36**	−.19	−.10	−.22	−.17	.83***	1.00
27. Instructional Support (FU)	−.06	.34	.32	.07	.02	−.22	−.04	.44***	.21	−.09	−.11	−.13	−.02	.20	−.08	−.29*	.09	.11	.06	−.05	−.25	−.18	−.29*	.05	.41***	.30***

Note. ECERS = Early Childhood Environment Rating Scale–Revised; BL = baseline; FU = follow-up; FRL = free/reduced lunch; LOTE = language other than English spoken at home; IEP = individualized education plan.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

TABLE 4

Primary Classroom-Level Impacts on Math Teaching Practices in the Spring

Outcome	*N*	Explore, adjusted *M* (*SD*)	Non-Explore, adjusted *M* (*SD*)	Difference (impact)	Effect size
Count of teacher-led math activities	95	2.53 (0.76)	0.98 (0.057)	1.55***	2.73
Minutes of teacher-led math activities	95	19.67 (6.62)	6.09 (7.65)	13.26***	1.78
Classrooms with at least one observed teacher-led math activity	95	0.71 (0.18)	0.34 (0.24)	0.37***	1.58
Dichotomous moderate math quality scores[a]	95	0.41 (0.18)	0.23 (0.19)	0.18***	0.95
Average math activity quality score if observed[b]	51	2.32 (0.27)	2.24 (0.31)	0.08	0.28
Mathematical Classroom Culture	95	2.93 (0.22)	2.50 (0.20)	0.43***	2.15

Note. Effect size is calculated by dividing the impact of the program (the difference between the means for the program group and the control group) by the standard deviation for the control group.

[a] Category is in contrast to classrooms with a low-quality score or no math activity observed. For each teacher-led math activity observed, quality was calculated by averaging across six items rated on a scale from 1 (*low*) to 5 (*high*). The scale assesses the extent to which the teacher explains the math concept underlying an activity, asks open-ended questions, and builds on children's answers, ideas, and strategies to extend their mathematical thinking. Scores at or above 2 were classified as having moderate to high quality. Classrooms were coded 0, or low-quality math instruction, if they had an average math quality rating below 2 or no quality rating. [b] For classrooms where a teacher-led math activity was observed, the average math activity quality score is calculated by averaging across nine items and then averaging across math activities for the final score; the score ranges from 1 (*strongly disagree*) to 5 (*strongly agree*), and assesses the extent to which teachers expanded children's conceptual understanding of math and extended children's mathematical thinking. This does not represent a true impact since the number of classrooms where at least one teacher-led math activity was observed was different between program and control groups (71% vs. 34%).

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

TABLE 5

Classroom-Level Impacts on the Number of Teacher-Led Math Activities and Informal Math Activities in Different Math Content Areas in the Spring

Math content area	*N*	Explore, adjusted *M* (SD)	Non-Explore, adjusted *M* (SD)	Difference (impact)	Effect size
Numbers	95	1.71 (0.49)	1.19 (0.46)	0.53***	1.15
Operations	95	0.28 (0.25)	0.04 (0.17)	0.24***	1.38
Geometry	95	0.35 (0.36)	0.13 (0.26)	0.22***	0.84
Spatial	95	0.04 (0.11)	0.12 (0.12)	−0.08***	−1.0
Patterning	95	0.16 (0.17)	0.05 (0.17)	0.10***	0.64
Measuring	95	0.15 (0.21)	0.03 (0.15)	0.11***	0.80

Note. Effect size is calculated by dividing the impact of the program (the difference between the means for the program group and the control group) by the standard deviation for the control group.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

reach the level of implementation seen in an effectiveness trial within the same school system at scale.

Nonetheless, the impacts' size is impressive, considering the Explore track had fewer workshops, fewer coaching sessions, and lower attendance rates than other studies of BB. For instance, participating teachers in the MPC project received 2 more training days: 6 days of training in MPC than 4 days in this study. The dosage for in-classroom coaching was much more substantial in the MPC project (3-hour, in-classroom coaching every week) compared to this study (1-hour, in-classroom coaching once a month). Moreover, teacher attendance at the Explore PD sessions was low: Only 18% of Explore sites had the expected number of teachers attending all four PL sessions.[4] This attendance rate was much lower than prior efficacy studies of BB; in the MPC study, teacher attendance at training sessions was high (87%, on average). Such less than ideal PD attendance is similar to what has been seen in other studies and, perhaps, unsurprising for a PD program at scale (Piasta et al., 2017).

The results (including the robustness checks) suggest that Explore's impact on the quality of instruction was mixed. Although Explore sites provided slightly higher quality math instruction than teachers in non-Explore sites, the differences were not always statistically significant. The lack of consistent statistically significant impacts on the quality scores suggests that the difference in observed math quality

TABLE 6
Interaction Results

Interaction	Count of teacher-led math activities	Minutes of teacher-led math activities	Classrooms with at least one teacher-led math activity (%)	Average math quality scores	Dichotomous moderate math quality scores	Mathematical Classroom Culture
Emotional Support (baseline) × Intervention	1.18 (0.84)	3.57 (8.18)	0.14 (0.23)	−0.12 (0.43)	0.22 (0.23)	0.83 (0.30)***
Classroom Organization (baseline) × Intervention	0.28 (1.34)***	8.61 (7.72)	0.33 (0.22)	0.33 (0.44)	0.45 (0.21)	0.66 (0.29)
Instructional Support (baseline) × Intervention	0.00 (0.48)*	1.22 (4.66)	0.10 (0.46)	−0.18 (0.25)	0.14 (0.13)	0.12 (0.18)

Note All of these analyses control for baseline classroom quality, child demographics, child gender, % of children who receive free/reduced-price lunch, % of children with an individualized education plan, % of children from homes that speak a language other than English, borough, and site type.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

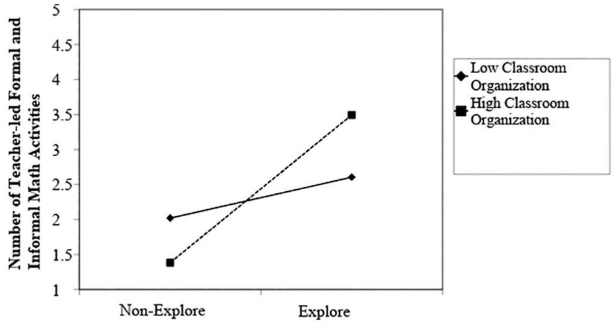


FIGURE 3. *Interaction effects for the number of teacher-led math activities and baseline classroom organization.*

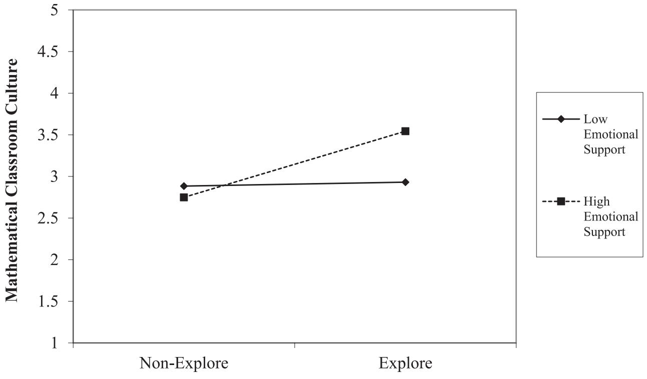


FIGURE 4. *Interactions effects for the mathematical classroom culture and baseline emotional support.*

may be driven by the difference in the presence of math instruction across the two groups of classrooms. However, in both groups, the degree to which teachers consistently used high-quality instructional strategies during math activities was relatively low overall—below a rating of a 2—meaning that teachers employed these strategies only some of the time.

It is likely that these limited effects on quality are driven by the fact that improving early math instruction can be difficult for teachers (Lee & Ginsberg, 2009); it requires teachers to know the content, understand children's thinking, engage in pedagogical practices that support learning, and see themselves as capable math teachers (Lee et al., 2009). Most PD studies do not assess impacts until the second year of implementation to allow teachers to learn and immerse themselves in the first year's curriculum (e.g., Morris et al., 2014). This article reports the results after the first year of a 2-year intervention; thus, we suspect we might not see impacts on math quality until the end of the second year of implementation.

Nonetheless, Explore classrooms had a positive and statistically significant impact on the classroom culture score, namely, the teachers' general mathematics education approach. Explore's impact on the classroom culture score suggests that the program successfully altered teachers' beliefs and dispositions beyond specific curriculum practices. Furthermore, a previous study of BB found that mathematical classroom culture mediated the relationship between the intervention's receipt with child outcomes (Clements et al., 2011). This is consistent with the literature supporting the connection between academic performance and general classroom features, including signs of mathematical activity, teachers who are knowledgeable and enthusiastic about mathematics, and teachers who frequently interact and respond to children (Clarke & Clarke, 2004; Clements & Sarama, 2007).

The robustness results (see Appendix B) build further confidence that any observed baseline nonequivalence in demographic composition and baseline quality measure of Explore and non-Explore sites is unlikely to be biasing the estimated impacts of Explore on teacher-led math practices as reported. Analyzing Explore's impact within a public school subsample had no appreciable effect on the observed findings' pattern, magnitude, or significance. When we subset the sites with overlapping preferences sets and those ranked Explore as "1," we found comparable results for the amount of math instruction but mixed results on the quality of math instruction.

Classroom Quality as a Moderator

Surprisingly, baseline classroom instructional support did not moderate the relation between the Explore and the amount and quality of math. That is, COEMET assesses the degree to which teachers use such instructional strategies as (1) asking open-ended questions, (2) formally extending children's math learning, and (3) explaining the math concept during activities—all strategies that encompass the CLASS instructional support domain. We suspect this could be due to instructional support in this sample, as in other samples across the United States, being relatively low (Burchinal, 2018).

The degree to which Explore affected the count of math activities depended on the level of classroom management before Explore. Implementing Explore, regardless of sites' baseline classroom management, increased the number of math activities compared to non-Explore sites. However, sites with high classroom management were able to conduct more math activities. This suggests that, perhaps, sites with higher classroom management skills before Explore were better equipped to implement. That is, teachers in well-organized classrooms that provide a structured environment where expectations and routines were delineated were able to implement more math activities (Bulotsky-Shearer et al., 2014). The BB curriculum is structured around weekly lesson plans consisting of four main instructional components: Whole Group, Small Group, Hands-On Math Centers, and Computer (Clements & Sarama, 2007). A higher degree of classroom organization, likely, facilitated teachers' ability to set up and implement multiple BB components throughout the preschool day.

Receiving the Explore PD was significantly related to higher mathematical classroom culture among sites with higher baseline emotional support. The mathematical classroom culture included items about the environment (e.g., math signs), teacher-child interactions (e.g., teachers actively interacting), and personal attributes of the teacher (e.g., showed curiosity about math ideas). Potentially, teachers who were more attuned to their children's needs were able to create a classroom that made math fun and to engage children in math activities positively. Improving the mathematical classroom culture is vital because it represents not adherence to the curriculum but rather the development of an environment that infuses math at every opportunity (Clements & Sarama, 2007). Implications for these moderation findings suggest that teachers with higher emotional support and classroom organization are better equipped to implement a math-specific PD program.

Limitations

Though this study is marked by numerous strengths, including the observations of classroom quality, rigorous design, and focus on a program at scale, there are several limitations. Most important, random assignment was not used as the method for allocating sites to tracks. Because we are relying on a natural experiment, we conducted several robustness checks to adjust for selection bias, but there is still the potential that our results are not internally valid. Similarly, this study's results may have limited generalizability to the broader set of sites across NYC's PKA system. Finally, we conducted our data collection, analyses, and made inferences at the site level. Due to sites being assigned to PD track at the site level and internal DECE-DOE processes for collecting data, we could not account for the multilevel nature of the sites, which have teachers and classrooms clustered within sites. This includes that we have observational data from only one classroom per site, which could lead to an over- or underestimation of effects.

Conclusion

Our results provide further evidence of the ability of a comprehensive PD program to improve teachers' implementation of math practices—despite the dearth of math instruction and preschool teachers' reported fear of math (e.g., Lee & Ginsberg, 2009). Our findings parallel other PD studies at scale (albeit mostly language- and literacy-focused PD) concerning the dosage of PD offered and impacts on teacher practice (e.g., Piasta et al., 2017; Weiland & Yoshikawa, 2013). Since child outcomes were not measured in this study, we do not know whether the Explore PD made a substantial enough impact on teacher outcomes to yield effects at the child level. Nonetheless, when interpreted within the extant literature, our findings raise some important questions regarding the field's approach to PD that have yet to be addressed. Specifically, the current PD system was designed to reflect recommendations for effective PD at scale (Hamre et al., 2017), yet our findings and others in the PD literature suggest that adhering to these general principles may not be sufficient to improve the quality of teacher practices (e.g., Pisata et al., 2017). More research is needed to understand the necessary infrastructure, resources, and other supports to improve the quality of teacher practice.

Appendix A

Priority Groups for Assigning Sites to PD Tracks

Priority group	How site ranked PD track	Recommended for PD track
1st priority	1st choice	Yes
2nd priority	2nd choice	Yes
3rd priority	3rd choice	Yes
4th priority	1st choice	No
5th priority	2nd choice	No
6th priority	3rd choice	No

Note. PD = professional development.

Appendix B

Robustness Checks

Method. A series of robustness checks are run to account for the potential of bias that may undermine the validity of the natural experiment (i.e., the sample was selected in a way that resulted in nonequivalent treatment and control groups). First, we examine sites that are similar in other ways outside the control of the treatment or predate the treatment. For instance, site type may be critical both to the choice of PD track (previous studies indicated that preschools within district schools emphasize academics over social-emotional or creative arts; Goodson & Moss, 1992; Phillips et al., 1994) and other characteristics of teachers and children. By constraining the sample to more similar sites across treatment and control groups, any observed and unobserved differences due to the design or chance may be minimized, allowing us to test the robustness of Explore's observed effects. Second, we reran our analyses to address the fact that how a site ranked their preferences for different PD tracks may reflect something about a site's approach or pedagogical mission, and that treatment sites chose to be switch to the Explore condition (but we do not know whether the sites in the control condition would have switched to Explore). To address this potential flaw, we examined the impacts on sites with similar preferences sets as well as just the subset of sites that ranked Explore as their first choice.

Robustness Checks of Differences in Baseline Characteristics. There were a few nonsignificant, but still notable, differences in the characteristics of the sites in the Explore and non-Explore sites, which raised potential concerns that any observed differences in teachers' practices might reflect these differences and not the implementation of Explore. Thus, to build further confidence in the interpretation of the impact results, we reestimated our models, examining impacts on outcomes, for the following subset of sites: (1) sites that ranked Explore as their first choice, (2) sites with similar preference choice sets, and (3) district schools only.

Our first and second set of robustness checks were conducted to account for the fact that site preference was a primary factor in how a site was assigned to track. It is important to try to minimize selection bias; how sites rank PD tracks was not random, and the factors that affect how sites rank PD tracks may likely be related to their future outcomes. Every PD track had unique rules that governed how site assignment to a PD track was granted. For instance, if sites that ranked Explore as "1" and Create as "2" were more likely to end up in the non-Explore sample than sites that ranked Explore "2" and Create "1," we might be worried that they differ on their initial buy-in and, ultimately, in their implementation of Explore. Alternatively, if the sites who were ex ante the most likely to have higher quality classrooms prefer a certain PD track, then it is difficult to untangle the effect of the intervention itself on program quality. Finally, and most important, by examining impacts among a subset of sites that had similar preference choice sets, we attempt to account for the fact that our treatment group is made up of Explore sites that chose to be reassigned to Explore, but we do not know whether our control group is made up of sites that would have also chosen to be reassigned to Explore.

Thus, impacts were examined again but only on among the subset of sites that had similar preference choice sets (i.e., Explore 1, Create 2; Explore 2, Create 1; see Table B1). This reduced our sample from 95 sites to 83 sites. There was only one baseline difference in site characteristics within this subset of sites; Explore sites were statistically significantly more likely to serve Black children (but notably, differences across other baseline characteristics were reduced). The impacts are shown in the first set of columns in Table 5 and show that the pattern, magnitude, and statistical significance of the impacts on teacher's math practices remain roughly the same. That is, positive impacts of Explore were found on the number of teacher-led minutes of math instruction, the number of math activities conducted by teachers, the quality of math instruction (both percentage of classrooms with moderate math quality scores and average math quality score), and the mathematical classroom culture score.

TABLE B1

Baseline Equivalence of the Sensitivity Analysis

	District schools–only analysis		Overlapping site ranking		Rated Explore No. 1	
Characteristics	Explore	Non-Explore	Explore	Non-Explore	Explore	Non-Explore
Site type						
Public schools (%)	—	—	41	25	54	29**
DOE-NYCEEC (%)	—	—	49	63	37	61†
ACS-NYCEEC (%)	—	—	5	11	3	10
PreK center (%)	—	—	5	0	6	0
Child demographics						
Race (%)						
Asian	14	12	18	22	15	25
Black	38	28	40	20***	37	25
Hispanic	38	48	28	36	35	32
White	8	10	11	19†	10	15
Multirace	2	2	3	3	3	3
Female	50	53	51	51	50	52
Poverty %	80	82	55	51	61	54
% with IEP	5	6	9	6	8	7
% LOTE	36	36	29	35	32	35
Program quality						
ECERS	3.53	4.03	3.97	4.28†	3.67	4.28†
CLASS Emotional Support	6.27	6.43	6.28	6.25	6.21	6.30
CLASS Classroom Organization	6.19	6.28	6.13	6.08	6.09	6.14
CLASS Instructional Support	2.88	3.88†	3.12	3.39	3.16	3.32
Joint test of all baseline characteristics	$F(16, 15) = 1.43, p = .92$		$F(19, 62) = 1.71, p = .06$		$F(19, 45) = 1.00, p = .48$	
Sample size						
Sites	21	11	39	44	35	31

Note. Effect size is calculated by dividing the impact of the program (the difference between the means for the program group and the control group) by the standard deviation for the control group. DOE-NYCEEC = Department of Education–New York City Early Education Center; ACS-NYCEEC = Administration for Children's Services–New York City Early Education Center; LOTE = language other than English spoken at home; IEP = individualized education plan; ECERS = Early Childhood Environment Rating Scale–Revised; CLASS = Classroom Assessment Scoring System.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

The second set of robustness checks subset the sample to only sites that ranked Explore as their first choice. This reduced our sample from 95 sites to 66 sites. In this subset of sites, Explore sites were statistically significantly more likely to be a public school setting (see Table B1). The results of the impact analysis are shown in the middle set of columns in Table B2. The findings on this smaller sample estimated positive impacts of Explore on the number of minutes of teacher-led math instruction, the number of teacher-led math activities conducted by teachers, and mathematical classroom culture. Compared to the full sample findings, the size and magnitude of impacts were similar, with the one exception being a lack of statistically significant impact on the measures of math quality.

Finally, as shown in Table B1 (district schools–only analysis), when the sample was limited to only district school sites, there were no baseline differences between Explore and non-Explore sites except for ECERS scores, a rating of classroom quality. As with the larger sample, non-Explore sites tended to be rated higher on ECERS-R. The higher ECERS scores in comparison sites versus treatment classrooms should bias estimates of treatment impact downward (as was the case for the full sample). The adjusted results are presented in Table B2 (district schools–only analysis). Explore sites in public schools had positive and statistically significant impacts on the number of teacher-led mat activities, the amount of time spent on teacher-led math activities, and mathematical classroom culture scores compared to non-Explore sites. These impacts were similar in the size and magnitude of the results with the full sample. The impacts on quality were not consistent with the full sample results: Positive and statistically significant effects were found for the average math activity quality score but not for classrooms with moderate math activity quality scores. The inconsistent finding for the quality scores may be due to the small sample size.

TABLE B2

Sensitivity Analyses for Impacts on Teacher Practices

		District school sample only				Overlapping site rankings				Rated Explore No. 1			
Outcome	*N*	Explore adjusted *M* (*SD*)	Non-Explore adjusted *M* (*SD*)	Difference (impact)	Effect size	Explore adjusted *M* (*SD*)	Non-Explore Adjusted *M* (*SD*)	Difference (impact)	Effect size	Explore Adjusted *M* (*SD*)	Non-Explore Adjusted *M* (*SD*)	Difference (impact)	Effect size
Count of teacher-led math activities	95	2.62 (1.85)	0.73 (1.13)	1.98**	1.67	2.74 (0.77)	0.98 (0.63)	1.76***	2.81	2.59 (1.14)	1.00 (0.86)	1.59***	1.85
Minutes of teacher-led math activities	95	19.73 (12.63)	1.80 (8.03)	17.93***	2.23	22.26 (6.55)	6.09 (8.00)	16.17***	2.02	20.26 (7.73)	7.70 (9.34)	12.56***	1.35
Classrooms with at least one teacher-led math quality score (%)	95	0.71 (0.32)	0.18 (0.33)	0.53***	1.62	0.74 (0.17)	0.34 (0.27)	0.40***	1.48	0.74 (0.19)	0.35 (0.25)	0.38***	1.55
Dichotomous moderate math quality scores[a]	95	0.29 (0.39)	0.18 (0.30)	0.10	0.34	0.42 (0.18)	0.23 (0.16)	0.19***	1.18	0.32 (0.18)	0.29 (0.17)	0.03	0.19
Average math activity quality score if teacher-led[b]	51	1.60 (0.82)	0.53 (0.89)	1.08***	1.21	2.37 (0.32)	2.11 (0.27)	0.26***	0.96	2.21 (0.28)	2.21 (0.30)	0.00	0.00
Mathematical Classroom Culture	95	2.86 (0.57)	2.38 (0.55)	0.48*	0.86	2.96 (0.30)	2.50 (0.23)	0.47***	1.99	2.92 (0.33)	2.54 (0.24)	0.38***	1.60
Sample size		21	11			39	44			35	41		

Note. Effect size is calculated by dividing the impact of the program (the difference between the means for the program group and the control group) by the standard deviation for the control group.

[a] Category is in contrast to classrooms with a low-quality score or no math activity observed. For each teacher-led math activity observed, quality was calculated by averaging across six items rated on a scale from 1 (*low*) to 5 (*high*). The scale assesses the extent to which the teacher explains the math concept underlying an activity, asks open-ended questions, and builds on children's answers, ideas, and strategies to extend their mathematical thinking. Scores at or above 2 were classified as having moderate to high quality. Classrooms were coded 0, or low-quality math instruction, if they had an average math quality rating below two or no quality rating. [b] For classrooms where a teacher-led math activity was observed, the average math activity quality score is calculated by averaging across nine items and then averaging across math activities for the final score; the score ranges from 1 (*strongly disagree*) to 5 (*strongly agree*), and assesses the extent to which teachers expanded children's conceptual understanding of math and extended children's mathematical thinking. This does not represent a true impact since the number of classrooms where at least one teacher-led math activity was observed was different between program and control groups (71% vs. 34%).

$^{\dagger}p < .10.$ $^{*}p < .05.$ $^{**}p < .01.$ $^{***}p < .001.$

Acknowledgments

Thank you to all to the following colleagues for your support collecting data: Damaris Rothe, Natalie Spitzer, and Parisa Spitaleri. Finally, we would like to thank the children, teachers, and directors who participated and made this research possible.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by grants from the Spencer Foundation.

Notes

1. The Units are available for programs to implement across all tracks and may be implemented in the control and treatment sites; thus, we examine the Explore's impact on the practices we would expect the Explore PD track to change uniquely–math practices.

2. Because of additional qualifications based on site characteristics that determined whether a site was assigned to the Thrive PD track, we do not include Thrive in this study.

3. In pre-K, NYCEECs do not collect lunch forms and do not report on free or reduced-price lunch status. In pre-K, students are not designated as English Language Learners, and they are screened for IEPs only at the parents' request.

4. The attendance rate for Explore sites was similar to rates of non-Explore sites (18% vs. 16%).

References

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science*, *333*(6045), 975–978. https://doi.org/10.1126/science.1204534

Barnett, W. S., Carolan, M. E., Squires, J. H., Brown, K. C., & Horowitz, M. (2017). *The state of preschool 2016: State preschool yearbook*. National Institute for Early Education Research.

Belfield, C. R., Nores, M., Barnett, S., & Schweinhart, L. (2006). The High/Scope Perry Preschool Program cost–benefit analysis using data from the age-40 followup. *Journal of Human Resources*, *41*(1), 162–190. https://doi.org/10.3368/jhr.XLI.1.162

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., Blair, C., Nelson, K. E., & Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child Development*, *79*(6), 1802–1817. https://doi.org/10.1111/j.1467-8624.2008.01227.x

Bloom, H. S., & Weiland, C. (2015). *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study* (Working paper). MDRC. https://doi.org/10.2139/ssrn.2594430

Bulotsky-Shearer, R. J., Bell, E. R., Carter, T. M., & Dietrich, S. L. (2014). Peer play interactions and learning for low-income preschool children: The moderating role of classroom quality. *Early Education and Development*, *25*(6), 815–840. https://doi.org/10.1080/10409289.2014.864214

Burchinal, M. (2018). Measuring early care and education quality. *Child Development Perspectives*, *12*(1), 3–9. https://doi.org/10.1111/cdep.12260

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., Barbarin, O. A., Sparling, J. J., & Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: An Abecedarian Project follow-up. *Developmental Psychology*, *48*(4), 1033–1043. https://doi.org/10.1037/a0026644

Clarke, D. M., & Clarke, B. A. (2004). Mathematics teaching in K-2: Painting a picture of challenging, supportive and effective classrooms. In R. Rubenstein, & G. Bright (Eds.), *Perspectives on teaching mathematics: 66th yearbook* (pp. 67–81). National Council of Teachers of Mathematics.

Clements, D. H., & members of the Conference Working Group. (2004). Part one: Major themes and recommendations. In D. H. Clements, J. Sarama, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 1–72). Lawrence Erlbaum.

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*(2), 136–163. https://doi.org/10.5951/jresematheduc.13.2.0136

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*(2), 443–494. https://doi.org/10.3102/0002831207312908

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, *333*(6045), 968–970.

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127–166. https://doi.org/10.5951/jresematheduc.42.2.0127

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812–850. https://doi.org/10.3102/0002831212469270

Cook, T. D., & Campbell, D. T. (1979). The design and conduct of true experiments and quasi-experiments in field settings. In In R. T. Mowday, & R. M. Steers (Eds.), *Research in organizations: Issues and controversies* (Reproduced in part). Goodyear.

Desimone, L. M., & Garet, M. S. (2015). Best practices in teacher's professional development in the United States. *Psychology, Society, & Education*, *7*(3), 252–263. https://doi.org/10.25115/psye.v7i3.515

Dobbie, W., & Fryer Jr, R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, *3*(3), 158–187. https://doi.org/10.1257/app.3.3.158

Dodge, K. A. (2009). Community intervention and public policy in the prevention of antisocial behavior. *Journal of Child Psychology and Psychiatry*, *50*(1–2), 194–200. https://doi.org/10.1111/j.1469-7610.2008.01985.x

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

Duncan, G. J., & Vandell, D. L. (2012). *Understanding variation in the impacts of human capital interventions on children and youth* (Irvine Network on Interventions in Development working paper). https://appam.confex.com/appam/2012/webprogram/Paper3857.html

Dunning, T. (2008). Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly*, *61*(2), 282–293. https://doi.org/10.1177/1065912907306470

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Early, D. M., Maxwell, K. L., Ponder, B. D., & Pan, Y. (2017). Improving teacher-child interactions: A randomized controlled trial of Making the Most of Classroom Interactions and My Teaching Partner professional development models. *Early Childhood Research Quarterly*, *38*, 57–70. https://doi.org/10.1016/j.ecresq.2016.08.005

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report*, *22*(1), 1–24. https://doi.org/10.1002/j.2379-3988.2008.tb00054.x

Goble, P., & Pianta, R. C. (2017). Teacher–child interactions in free choice and teacher-directed activity settings: Prediction to school readiness. *Early Education and Development*, *28*(8), 1035–1051. https://doi.org/10.1080/10409289.2017.1322449

Goodson, B., & Moss, M. (1992). *Analysis report: Predicting quality of early childhood classrooms: Observational study of early childhood programs*. Abt Associates.

Gormley, W. T., Jr., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*(6), 872–881. https://doi.org/10.1037/0012-1649.41.6.872

Gulamhussein, A. (2013). *Teaching the teachers: Effective professional development in an era of high stakes accountability*. Center for Public Education.

Hamre, B. K. (2014). Teachers' daily interactions with children: An essential ingredient in effective early childhood programs. *Child Development Perspectives*, *8*(4), 223–230. https://doi.org/10.1111/cdep.12090

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*(3), 1257–1274. https://doi.org/10.1111/cdev.12184

Hamre, B. K., Partee, A., & Mulcahy, C. (2017). Enhancing the impact of professional development in the context of preschool expansion. *AERA Open*, *3*(4), Advance online publication. https://doi.org/10.1177/2332858417733686

Harms, T., Clifford, R., & Cryer, D. (2003). *Early Childhood Environment Rating Scale–Revised*. Teachers College Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hoxby, C. M (2001). *How school choice affects the achievement of public school students* [Paper prepared for Koret Task Force meeting]. Hoover Institution, Stanford, CA.

Hustedt, J. T., Barnett, W. S., Jung, K., & Thomas, J. (2007). *The effects of the Arkansas Better Chance Program on young children's school readiness*. National Institute for Early Education Research. http://nieer.org/research-report/the-effects-of-the-arkansas-better-chance-program-on-young-childrens-school-readiness

Institute of Education Sciences & National Science Foundation. (2013). *Common guidelines for education research and development: A report from the Institute of Education Sciences, US Department of Education and the National Science Foundation*.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, *100*(1), 96–104. https://doi.org/10.1037/0022-0663.100.1.96

Lee, J. S., & Ginsburg, H. P. (2009). Early childhood teachers' misconceptions about mathematics education for young children in the United States. *Australasian Journal of Early Childhood*, *34*(4), 37–45. https://doi.org/10.1177/183693910903400406

Leithwood, K., & Jantzi, D. (2009). A review of empirical evidence about school size effects: A policy perspective. *Review of Educational Research*, *79*(1), 464–490. https://doi.org/10.3102/0034654308326158

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

Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., & Hofer, K. G. (2015). The prekindergarten age-cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis*, *37*(3), 296–313. https://doi.org/10.3102/0162373714547266

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, *26*(1), 33–51. https://doi.org/10.1016/j.econedurev.2005.09.008

Markussen-Brown, J., Juhl, C. B., Piasta, S. B., Bleses, D., Højen, A., & Justice, L. M. (2017). The effects of language-and literacy-focused professional development on early educators and children: A best-evidence meta-analysis. *Early Childhood Research Quarterly*, *38*, 97–115. https://doi.org/10.1016/j.ecresq.2016.07.002

Morris, P., Lloyd, C. M., Millenky, M., Leacock, N., Raver, C. C., & Bangser, M. (2013). *Using classroom management to improve preschoolers' social and emotional skills: Final impact and implementation findings from the Foundations of Learning Demonstration in Newark and Chicago*. MDRC. https://doi.org/10.2139/ssrn.2202401

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

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Phillips, D. A., Voran, M., Kisker, E., Howes, C., & Whitebook, M. (1994). Child care for children in poverty: Opportunity or inequity? *Child development*, *65*(2 Spec No), 472–492.

Pianta, R., Hamre, B., Downer, J., Burchinal, M., Williford, A., Locasale-Crouch, J., Howes, C., La Paro, K., & Scott-Little, C. (2017). Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. *Early Education and Development*, *28*(8), 956–975. https://doi.org/10.1080/10409289.2017.1319783

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3*. Paul H. Brookes.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*(4), 431–451. https://doi.org/10.1016/j.ecresq.2008.02.001

Piasta, S. B., Justice, L. M., O'Connell, A. A., Mauck, S. A., Weber-Mayrer, M., Schachter, R. E., Farley, K. S., & Spear, C. F. (2017). Effectiveness of large-scale, state-sponsored language and literacy professional development on early childhood educator outcomes. *Journal of Research on Educational Effectiveness*, *10*(2), 354–378. https://doi.org/10.1080/19345747.2016.1270378

Presser, A. L., Clements, M., Ginsburg, H., & Ertle, B. (2012). *Effects of a preschool and kindergarten mathematics curriculum: Big Math for Little Kids*. Center for Children and Technology. http://cct.edc.org/publications/effects-preschool-andkindergarten-mathematics-curriculum-big-math-little-kids-final

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. https://doi.org/10.3102/0162373707299460

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *1*(2), 89–119. https://doi.org/10.1080/19345740801941332

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*(3), 489–502. https://doi.org/10.1016/j.ecresq.2011.12.002

Unterman, R., Bloom, D., Byndloss, D., & Terwelp, E. (2016). *Going away to school: An evaluation of SEED DC*. MDRC.

Wasik, B. A., & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology*, *103*(2), 455–469. https://doi.org/10.1037/a0023067

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352–360. https://doi.org/10.3102/0013189X14553660

Weiland, C., McCormick, M., Mattera, S., Maier, M., & Morris, P. (2018). Preschool curricula and professional development features for getting to high-quality implementation at scale: A comparative review across five trials. *AERA Open*, *4*(1), 2332858418757735. https://doi.org/10.1177/2332858418757735

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112–2130. https://doi.org/10.1111/cdev.12099

Whitehurst, G. J., & Chingos, M. M. (2011). *Class size: What research says and what it means for state policy*. Brookings Institution.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122–154. https://doi.org/10.1002/pam.20310

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Magnuson, K. A., Phillips, D., & Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. Foundation for Child Development.

Zaslow, M., Tout, K., Halle, T., Whittaker, J. V., & Lavelle, B. (2010). *Toward the identification of features of effective professional development for early childhood educators. Literature review*. Office of Planning, Evaluation and Policy Development, U.S. Department of Education.

Authors

NATALIA M. ROJAS is a National Science Foundation postdoctoral research fellow at New York University School of Medicine. She studies system-level supports for classroom quality with a specific interest in dual language learners.

PAMELA MORRIS is a professor of applied psychology at New York University's Steinhardt School of Culture, Education, and Human Development. Having spent a decade in policy research at MDRC before joining the faculty at New York University, Dr. Morris has spent two decades working at the intersection of social policy, practice, developmental psychology, and education.

AMUDHA BALARAMAN is the director of research and evaluation at New York University Department of Education. She is interested in increasing school readiness among children.