NBER WORKING PAPER SERIES

LOTTERY-BASED EVALUATIONS OF EARLY EDUCATION PROGRAMS: OPPORTUNITIES AND CHALLENGES FOR BUILDING THE NEXT GENERATION OF EVIDENCE

Christina Weiland
Rebecca Unterman
Susan Dynarski
Rachel Abenavoli
Howard Bloom
Breno Braga
Ann-Marie Faria
Erica H. Greenberg
Brian Jacob
Jane Arnold Lincove
Karen Manship
Meghan McCormick
Luke Miratrix
Tomás E. Monarrez
Pamela Morris-Perez
Anna Shapiro
Jon Valant
Lindsay Weixler

Working Paper 30970
http://www.nber.org/papers/w30970

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2023

After the third author, authors are listed alphabetically. The Spencer Foundation provided funding for this work through their conference grants program. The authors would like to thank the funders of the site projects as described: Institute of Education Sciences (Boston, DC, New Orleans, Montessori, and New York teams); Arnold Ventures (Boston); and the Heising-Simons Foundation (DC). Correspondence should be addressed to Christina Weiland at weilandc@umich.edu. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Lottery-Based Evaluations of Early Education Programs: Opportunities and Challenges for Building the Next Generation of Evidence

Christina Weiland, Rebecca Unterman, Susan Dynarski, Rachel Abenavoli, Howard Bloom, Breno Braga, Ann-Marie Faria, Erica H. Greenberg, Brian Jacob, Jane Arnold Lincove, Karen Manship, Meghan McCormick, Luke Miratrix, Tomás E. Monarrez, Pamela Morris-Perez, Anna Shapiro, Jon Valant, and Lindsay Weixler

NBER Working Paper No. 30970
February 2023
JEL No. I20,I21

ABSTRACT

Lottery-based identification strategies offer potential for generating the next generation of evidence on U.S. early education programs. Our collaborative network of five research teams applying this design in early education and methods experts has identified six challenges that need to be carefully considered in this next context: 1) available baseline covariates may not be very rich; 2) limited data on the counterfactual; 3) limited and inconsistent outcome data; 4) weakened internal validity due to attrition; 5) constrained external validity due to who competes for oversubscribed programs; and 6) difficulties answering site-level questions with child-level randomization. We offer potential solutions to these six challenges and concrete recommendations for the design of future lottery-based early education studies.

Christina Weiland
University of Michigan
School of Education
610 E. University Ave
Ann Arbor, MI 48104
weilandc@umich.edu

Rebecca Unterman
MDRC
200 Vesey Street 23rd Floor
New York, NY 10281-2103
Rebecca.Unterman@mdrc.org

Susan Dynarski
Harvard University Graduate
School of Education
13 Appian Way
Cambridge, MA 02138
and NBER
susan_dynarski@harvard.edu

Rachel Abenavoli
New York University
82 Washington Square East
New York, NY 10003
rachel.abenavoli@nyu.edu

Howard Bloom
MDRC
200 Vesey Street 23rd Floor
New York, NY 10281-2103
howard.bloom@mdrc.org

Breno Braga
Urban Institute
500 L'Enfant Plaza SW
Washington, DC 20024
bbraga@urban.org

Ann-Marie Faria
American Institutes for Research
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
afaria@air.org

Erica H. Greenberg
Urban Institute
500 L'Enfant Plaza SW
Washington, DC 20024
egreenberg@urban.org

Brian Jacob
Gerald R. Ford School of Public Policy
University of Michigan
735 South State Street
Ann Arbor, MI 48109
and NBER
bajacob@umich.edu

Jane Arnold Lincove
UMBC School of Public Policy
1000 Hilltop Cir
Baltimore, MD 21250
jlincove@umbc.edu

Karen Manship
American Institutes for Research
100 Europa Dr., Suite 315
Chapel Hill, NC 27517
kmanship@air.org

Meghan McCormick
MDRC
200 Vesey Street 23rd Floor
New York, NY 10281-2103
Meghan.McCormick@mdrc.org

Luke Miratrix
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA 02138
luke_miratrix@gse.harvard.edu

Tomás E. Monarrez
Urban Institute
500 L'Enfant Plaza SW
Washington, DC 20024
TMonarrez@urban.org

Pamela Morris-Perez
New York University
82 Washington Square East
New York, NY 10003
pam7@nyu.edu

Anna Shapiro
University of Virginia
405 Emmet St S.
Charlottesville, VA 22904
aks6q@virginia.edu

Jon Valant
Brookings Institution
1775 Massachusetts Ave NW
Washington, DC 20036
JValant@brookings.edu

Lindsay Weixler
Tulane University
6823 St Charles Ave
New Orleans, LA 70118
lweixler@tulane.edu

A data appendix is available at http://www.nber.org/data-appendix/w30970

Lottery-Based Evaluations of Early Education Programs:

Opportunities and Challenges for Building the Next Generation of Evidence

I. Introduction

Decades of research show that preschool[1] helps prepare children for kindergarten, and in some contexts, improves participants' outcomes into adulthood (Phillips et al., 2017; Yoshikawa et al., 2013). But much of this evidence comes from older, small programs, which differ substantially from modern preschools in their curriculum, funding, diversity of children served, and alternative options. More evidence on today's large-scale public programs is needed for guiding policy and practice (Phillips et al., 2017). This is especially the case in the wake of a pandemic that has been particularly devastating for the early care and education sector (Weiland et al. 2021) and in light of policy proposals to expand public preschool program to all U.S. three and four-year-olds (White House, 2021).

Lottery-based school assignment systems used in many cities across the U.S. have potential for helping to generate this needed evidence. In these systems, when programs are over-subscribed, a random process is used to choose among the applicants. Sometimes, these lotteries are generated via separate applications to individual schools and other times, by centralized school choice systems across an entire district. In both cases, this creates a natural experiment in which some children are granted access to particular schools or programs and others are not.

In the elementary and secondary school contexts, researchers have leveraged this random assignment to estimate the causal impacts of charter schools (Abdulkadiroğlu et al., 2011;

[1] We use the term "preschool" to refer to center-based care and education settings for three to five year olds. We also use "Pre-K' and "prekindergarten" when discussing specific programs that self-label using those terms.

Dynarski, Hubbard, Jacob, & Robles, 2019; Unterman, Bloom, Byndloss, & Terwelp, 2016; Unterman, 2017) and small schools of choice (Bloom & Unterman, 2014). This design has been leveraged in preschool in only two peer-reviewed studies to date (Gray-Lobe, Pathak, & Walters, 2023; Weiland et al., 2020), though at least five teams total (represented on our authorship team) are now leveraging this methodological approach to investigate policy and practice questions in large-scale systems.

In this paper, we bring together lessons and examples drawn from a collaborative network comprised of these five teams and a group of methods experts that illustrate how, when moving the lottery design into a new context, there are shared challenges that need to be carefully considered. The six challenges we cover are: 1) available baseline covariates may not be very rich; 2) limited data on the counterfactual; 3) limited and inconsistent outcome data; 4) weakened internal validity due to attrition; 5) constrained external validity due to who competes for oversubscribed programs; and 6) difficulties answering site-level questions with child-level randomization.

As we illustrate, these challenges are not necessarily unique to early education studies but in many cases, are *exacerbated* compared to lottery-based studies with older children or to other causal designs with preschool programs. Following the example of pedagogical guides that have helped improve applied randomized trial and regression discontinuity studies in education (Calonico et al., 2017; Duflo, Glennerster, & Kremer, 2007; Imbens & Lemieux, 2008; Lipsey et al., 2015; Murnane & Willett, 2010), our primary goal is to improve the application of the lottery-based design in current and future early education studies. Our secondary goal is also to serve as a case study more broadly of how *context* can affect study design in applied education work. In our view, this study design has the potential to provide much-needed evidence on many

critical early education questions. But without careful attention to its particularities, we fear it instead could be a source of randomization in search of a question. In other words, the tail could wag the dog and the opportunity to systematically tackle the most pressing questions in the field will not be realized.

In the sections that follow, we first explain the design and describe potential opportunities for building new evidence on early education programs using lottery-based designs. Then, drawing from the experiences of our five teams, we detail and provide examples of six challenges and possible solutions in early education lottery studies that are critical to consider a priori. We conclude with recommendations for the design of future lottery-based early education studies.

I.A Basic features of lottery-based studies

Lottery-based studies of education programs are possible because of the school choice systems now in place in many U.S. localities. The design of these systems vary from place to place. For example, in some school settings families submit individual applications to individual schools that then conduct their own lotteries, when there are more applicants than seats. In these studies, students who won the lottery formed the treatment group and those who lost, the control group (e.g., Abdulkadiroğlu et al., 2011; Dynarski et al., 2019; Unterman et al., 2016; Unterman, 2017). Standard methods in randomized trials (e.g., Angrist & Pischke, 2008; Bloom, 2005; Murnane & Willett, 2010) were then used to estimate both the impacts of treatment assignment and, under assumptions, of enrollment.

Other studies have leveraged school choice systems that are based on the deferred acceptance algorithm (Abdulkadiroğlu, 2011; Roth, 2008). While the specific assignment rules vary from setting to setting, this approach allows applicants to centralized systems, such as a

large school district, to reveal their true preference order and reduce gaming behaviors, such as ranking a less desired and less popular school first to improve chances of a match. In these systems in place in many large U.S. school districts, parents rank schools within a given set of choices, and slots are assigned based on their preferences as much as possible. Schools can rank applicants according to particular criteria as well. For example, they can give higher preference – and thus a greater likelihood of a match – to students with siblings in the school already and/or students who live in a particular geographic area. Each family is assigned a random number (unknown to them) at the beginning of the process. As with individual lotteries, when programs are over-subscribed, the random number is used as a tie-breaker (or coin flip) between children with the same priority and preference for the school.[2]

There are multiple analytic approaches to estimating impacts of a given education program leveraging the lotteries created by the deferred acceptance algorithm. One is to leverage only students' first-choice lottery (Lincove, Valant, & Cowen, 2018; Weiland et al., 2020). Another is using the first lottery in which a student competes regardless of choice order (e.g., if the student was shut out of their first choice entirely but then competed in a lottery for their second choice; Bloom & Unterman, 2014). More recently, scholars have developed deferred acceptance (DA) propensity score or assignment score approaches with the goal of including more students in the sample, increasing the statistical power and the generalizability of the impact estimates (Abdulkadiroğlu, Angrist, Narita, & Pathak, 2017). Empirical work comparing the first choice and first lottery approach in New York City's Small School of Choice (Bloom & Unterman, 2014) and the Boston Prekindergarten program (Weiland et al., 2020)

[2] Note that although some systems assign a global lottery number to each child, others assign each child a different lottery number for every program to which they're applying.

found no meaningful differences in treatment impacts between the first lottery and first choice analytic approaches. A similar rigorous analysis comparing the estimates from these two approaches with the newer assignment score approach across a diverse set of sites would greatly add to the field's understanding of the tradeoffs of these approaches. We return to this in our recommendations section.

Regardless of their analytic approach, empirical studies show that the lotteries generated by these school choice systems have strong *internal validity* – i.e., they result in treatment and control groups that were essentially randomized in a coin-flip-like procedure and that are equal in expectation before a given intervention began (Murnane & Willet, 2010). However, importantly, not all applicants in these systems are randomized, no matter the analytic approach the research team chooses. Only students who compete for *oversubscribed* schools are randomized and sometimes, only a minority of students are randomized to a relatively small number of schools. This has implications for external validity, or the generalizability of impacts estimated using this approach – an important issue we return to in our challenges section.

I.B. Deferred acceptance lottery assignment example

To build intuition, in Figure 1, we provide a concrete idea of what the matching process looks like for a hypothetical preschool applicant in a deferred acceptance choice system, following the DC Public Schools explainer for parents (My School DC, 2019). In our example, not all preschool applicants are assigned a seat – i.e., the treatment-contrast is between the preschool program and all local alternatives to it – and the researcher wants to identify the effect of winning a seat in the program versus being lotteried out. The numbers on children's shirts are their random lottery numbers. As shown in Figure 1, a child's family has ranked three schools they would like her to attend – North, West, and East Elementary Schools. She is the only child

with glasses in Figure 1, with random number 16. For simplicity, we will refer to her as Student 16. Her first choice – North – gives priority to students with siblings. Her second and third choices give priority to both students with siblings and children with a geographic area preference that we refer to as in-boundary status. These priorities are hierarchical (e.g., the system assigns those with sibling and in-boundary-status first, then siblings, and then children with in-boundary status). Student 16 has no priority at North, sibling preference at school West, and in-boundary preference at East.

Every student in the system is assigned a random number (unknown to them) as the first step in the assignment process. Student 16's position in line reflects the combination of her priority at each school and her random number within her priority group. Her ranked schools can each admit 20 students total and by the time her application is considered, they have different numbers of seats still unfilled (assume as in DC, that the seats were filled by students who attended the school's three-year-old program in the prior year and are "moving up" to the four-year-old program). Student 16 is unmatched to her first choice (North) because she is ninth in line and only two seats are open. Her second choice (West) has five seats open and she is sixth in line so she is again unmatched. Her third choice (East) has 5 seats open; she is third in line and matches here.

In a first-choice lottery analytic approach, Student 16 would not be part of the lottery sample; her first choice (North) was filled before her priority group was considered. In a first-lottery analytic approach, Student 16 is in the control group for the intent to treat estimates of the effects of being randomly assigned to the preschool program at West (i.e., due to competing in a siblings lottery at West and not matching there) and a crossover or always-taker in a local average treatment effect analysis of the effects of enrolling in the preschool program (i.e., due to

her match at East Elementary). In an assignment score analytic approach, Student 16 is considered a member of the treatment group, with a probability of treatment assignment that falls between 0 and 1, since she faced risk of not being assigned to the program. In this regression-based analysis, she will be analyzed in an assignment score block with other students that had a similar probability of assignment.[3]

I.C. Data in lottery-based studies

Another important feature of lottery-based studies is the data used, beyond students' choice data. To our knowledge, all lottery-based studies conducted to date have relied solely on administrative data, or data collected as part of the typical operation of a given district or school system. Student characteristics like race/ethnicity, gender, and dual language learner status, for example, are commonly tracked in educational administrative data. Other commonly available fields in administrative data include students' past and future test scores, attendance, disciplinary records, special education status, and grade retention – i.e., potential outcomes in a lottery-based study. To date, researchers in lottery-based evaluations have not engaged in primary data collection such as surveys or classroom observations. However, due to more limited administrative data available for early education studies, several of our five teams are now attempting to collect such data, as we detail later in this article.

II. Advantages of lottery-based studies for answering pressing questions in early education

[3] Note that in the first choice and first lottery analytic approach, one assumption underlying the analysis of the effects of enrollment – that always-takers in both the treatment and control groups (i.e., children who would have enrolled in the preschool program regardless of their first choice treatment assignment status) experienced the same effect of enrollment – is difficult to evaluate (Weiland et al., 2020). In an assignment score analytic approach, always-takers are either removed from the analysis because they have probability of treatment assignment of either 0 or 1 or included as enrollees, if their probability falls between 0 and 1. This is one of the key differences in these analytic approaches where additional applied work would be useful in preschool research to understanding the tradeoffs of the approaches and implications for impact estimates.

Thousands of families now apply to public preschool programs that use lottery-based assignment systems, presenting opportunities to address new research needs with no or limited disruption to a locality's standard operations. Lottery-based studies too have potential strengths over alternatives. First, lottery studies offer the opportunity to study policy initiatives in real-time and in their natural form. It can take years otherwise to rally support for and design an experimental test that can identify the causal effects of an intervention or policy. When random assignment changes natural operations as well, there is a possibility too that any detected effects are lottery-induced (i.e., John Henry and Hawthorne effects; Murnane & Willettt, 2010) – a threat ruled out (or at least reduced) in naturally occurring lotteries.

In addition, in randomized trials, many families are reluctant to consent in studies or simply forget to return consent forms. This threatens external validity, as consenting families may not be representative of the population of interest. Notably, working with the constraints of their context and design, the consent rate in one of the directly assessed cohorts of the randomized trial of the Tennessee Voluntary Pre-K study was only 24% (Lipsey, Farran, & Durkin, 2018). And even in randomized trials with relatively high rates of parental consent – for example 80% or higher – researchers may still find some differences in the characteristics of students who consent and those in the broader population. There can also be biasing attrition among those who consent – a threat to internal validity.

The potentially large numbers of students randomly assigned in naturally occurring lotteries each year also may permit more precise estimation of effects for important subgroups, particularly if leveraged across multiple cohort years. For example, there is evidence that dual-language learners benefit more from public preschool programs than their monolingual peers (Phillips et al., 2017), but randomized trials and birthday-cutoff-based regression discontinuity

studies (Gormley, Phillips, & Gayer, 2008; Weiland & Yoshikawa, 2013) often lack the statistical power to examine effects separately by specific home language. If enough members of subgroups compete in naturally occurring lotteries, lottery-based studies may permit more specificity in estimating effects for different language subgroups which could be very informative, given stark differences in language structure in Spanish versus Vietnamese, immigration histories, and home cultures.

Another feature of the lottery-based design that may be beneficial in building the next generation of evidence is that random assignment occurs *within lottery blocks* – i.e., within smaller sets of applicants to particular schools. For example, returning to Figure 1, Student 16 did not compete against every student for West Elementary (her second choice and her first lottery); she only competed against those with the same preference to school (i.e., the students shown with orange shirts). Her block in a first choice and first lottery analytic approach then is the West Elementary and sibling preference combination. Essentially, each of these blocks represent "mini-experiments" within the full applicant sample. Recent advancements in evaluation methods have highlighted how blocked random assignment can be used to move beyond average impacts to examine how effects vary across schools and the factors that predict this variation (Bloom, Raudenbush, Weiss, & Porter, 2017). For example, in a Boston Prekindergarten lottery-based study, effects on all third-grade outcomes varied substantially across blocks and the best school-level predictor of this variation was school standardized test scores (Unterman & Weiland, 2020). Preschool lottery contexts are very promising for additional such evidence. Blocked random assignment otherwise can be quite difficult to implement and is often under-powered for impact variation analyses in the early education

context due to factors like small numbers of classrooms in centers compared to K-12 settings (Sabol et al., 2022).

There may too be a parallel need for new research designs in the changing policy context. For example, universal public preschool is a priority of the Biden administration (White House, 2021) and multiple states and cities have moved in recent years to fund their own universal programs. These changes mean that researchers will no longer be able to rely on the kinds of scarcity and over-subscription that have permitted past studies of the causal effects of a given public preschool program versus alternatives, since all children in those systems now will be offered a seat (e.g., Lipsey et al., 2018; Puma et al., 2012). The changing policy context also raises new policy questions and thus introduces a need for a new generation of early education evidence. For example, some localities have introduced public preschool programs in part to attract and retain students in a given system – a new outcome to the literature – and two lottery-based early education studies indeed have found large positive outcomes on this outcome (Monarrez, Greenberg, Luetmer, & Chien, 2020; Weiland et al., 2020). With the vast majority of three to five year olds already in out-of-home care of some kind, some scholars have argued too for a pivot away from preschool versus none questions to a focus on *how* to build high-quality programs at scale, such as through comparing types of ECE or features of ECE (Bassok & Engel, 2019; Weiland, 2018). Lottery-based methods may provide opportunities to meet the new moment and new needs in the field.

III. Five Current Lottery-based Early Education Studies

Before turning to the challenges that lottery-based design presents in the early education context, we briefly describe the aforementioned five ongoing lottery-based early education studies represented among our authorship team, which provides the basis for our understanding

of and sensitivity to these challenges. We also summarize key information about these five studies in Table 1. Together, the site-based teams and methods experts form a collaborative network aiming to identify best practices for design and analysis, common challenges, and potential solutions for this future preschool research. As we describe below, each of the five teams too are addressing pressing questions in the field and breaking new ground as part of the next generation of evidence on the impacts of public preschool programs.

Boston Instructional Alignment Study. Curriculum alignment has emerged as a leading hypothesis about how best to build on children's preschool gains, so that preschool attenders do not merely repeat again the same content in kindergarten that they have already learned and therefore lose the opportunity to build on their preschool skills (Harding, McCoy, & McCormick, 2020; Stein & Coburn, 2021). However, limited rigorous empirical work has examined the effects of alignment. Only two studies have done so using study designs that could identify causality, both focused on math curriculum alignment and both finding positive effects (Clements, Sarama, Wolfe, & Spitler, 2013; Mattera, Jacob, MacDowell, & Morris, 202).

Using naturally occurring lotteries from Boston's application of the deferred acceptance algorithm and in partnership with the Boston Public Schools Department of Early Childhood, the study team comprised of researchers at MDRC and the University of Michigan is examining the impact of Boston's rollout of an aligned prekindergarten and kindergarten curriculum and professional development approach on children's language, literacy, and math skills in third grade (McCormick et al., 2022). The study breaks new ground in the field as the first-ever test of a district-created aligned curriculum across multiple learning domains and of a district rollout approach in the early years. In addition, the study will examine a set of exploratory research questions that estimate impacts on school persistence, attendance, receipt of special education

services, and grade retention, as well as whether effects vary by student subgroup characteristics. The study team is leveraging administrative data on three cohorts of students who applied to the program in 2012-2013, 2013-2014, and 2014-2015 to estimate impacts, for a lottery sample total of 2,656 students (out of 10,318 applicants, or 26%). A complier average causal effect analysis will estimate the effect of a student winning their first lottery and enrolling in the aligned school, compared with students that lost their first lottery and did not enroll in an aligned school.

DC Public Prekindergarten: Impacts on three year olds. Policy proposals under both the Obama and Biden administrations aimed to expand public preschool to all three and four year olds in the country (White House, 2013; 2021). Although there is ample evidence that such programs improve the school readiness of four year olds (Phillips et al., 2017), there is very little such evidence for three year olds, particularly using experimental methods in large samples (Head Start and Early Head Start are the exception; Love et al., 2005; Puma et al., 2012). This is due to the very practical reason that only one U.S. locality – Washington DC – offers public preschool to all three year olds in the District.

Since 2019, a team of researchers from the Urban Institute have been studying DC's program with support from the DC Office of the State Superintendent of Education. Their work spans both retrospective impact analysis of recent cohorts of both three and four year olds, as well as a prospective study of the impacts on three year olds (in collaboration with researchers at the University of Michigan and School Readiness Consultants). As summarized in Table 1, the randomized subsample size for the retrospective study is approximately 5,600 students (about 22% of applicants to the three-year-old program), while for the prospective study of the 2023 and 2024 cohorts, the target sample size is 2,500. Outcomes drawn from administrative data are similar to those in the Boston study (with the additions of school and residential mobility

outcomes). Prospective study outcomes include directly assessed measures of children's language, literacy, math, executive function, social emotional skills, and racial attitudes at the end of their three-year-old and four-year-old years, with plans to follow children beyond these years in future work. The team is using the assignment score analysis strategy described earlier (Abdulkadiroğlu et al., 2017; Monarrez et al., 2020) to estimate the impacts of enrolling in the program versus being randomized out and experiencing a different care setting in the three-year-old year.

Montessori. There are currently over 3,000 Montessori schools in the U.S., 560 of which are public schools and over 150 of which serve public preschool and kindergarten students (National Center for Montessori in the Public Sector, n.d.). Despite the model's popularity and growing prevalence in public schools, no large-scale evaluation of the efficacy of the Montessori model on children's academic, social, and emotional skills has been conducted until now. This evidence is critical for addressing the question about *what kind* of public preschool can produce which learning gains and for which students.

A team of researchers at AIR are collaborating to conduct the study. Drawing on a sample of 22 public Montessori schools around the U.S. that use lotteries to admit 3-year-old students, the team aims to estimate the impacts of the Montessori model through the end of kindergarten. They also plan to explore heterogeneity by student subgroup, incorporation of Montessori principles (i.e., fidelity), and the counterfactual. Outcome measures will be directly assessed by trained study staff; these will include widely used measures of language, literacy, math, and executive function, as well as more novel measures that tap constructs that align directly with the Montessori theory of change around building persistence and problem solving skills. Unique among the five teams, their lotteries are drawn from both applications to

individual oversubscribed schools and to schools participating in centralized choice systems using the deferred acceptance algorithm.

New Orleans Study of PK Quality. High-quality PK programs can lead to substantial short-term academic and cognitive gains for children (e.g., Gormley et al., 2005; Wong et al., 2008). However, how best to define and measure quality in early childhood education remains an open question. Prior research on "high-quality" programs has used a variety of definitions, including both structural and process features of care (Yoshikawa et al., 2013). Notably, however, government and research-based definitions may not match parents' definitions of quality. Differences could arise if parents' quality criteria differ (e.g., if parents incorporate elementary school considerations when choosing school-based PKs) or if parents judge programs differently using similar criteria (e.g., parents have different ways of assessing teacher quality).

With this proposed study, a team of researchers at Tulane University, the University of Maryland, and the Brookings Institute will compare government-defined quality and parent-defined quality. The former draws on scores obtained through systematic classroom observations using the CLASS measure (Pianta & Hamre, 2009); the latter draws on parents' ranked requests. Using New Orleans' centralized school-assignment lottery, the team will examine how children and families' short-term academic, cognitive, and socio-emotional outcomes are affected by winning a seat in (a) their top-choice PK programs or (b) PK programs rated highly by state government. The team will use data from seven cohorts of applicants (2017-18 through 2023-24), with an estimated lotteried sample size of 4,500 (of roughly 15,000 total applicants, 30%) to calculate treatment effects. Exploratory analyses will examine the role of elementary school and teacher quality in sustaining gains, teachers' and administrators' beliefs about the effects of PK, and effects of offering PK on school composition and outcomes.

New York City Pre-K for All Professional Learning (PL) Study. Over the last 15 years, rigorous studies have shown that some kinds of preschool programs produce larger child learning gains than others (Phillips et al., 2017; Yoshikawa et al., 2013). These studies have pushed beyond the question of whether preschool works (or not) to *how* to deliver high-quality preschool experiences. For example, models that use play-based curricula with a scope and sequence and that focus on a particular learning domain outperform those that use more general curricula that do not share these features (Clements & Sarama, 2008; Clements et al., 2013; Morris et al., 2014). Such findings have helped to fuel policy and practice attention to the specific malleable, active ingredients in large, at-scale programs.

New York University researchers partnered with the New York City Department of Education to answer a pressing *how* question in their context – the effects of several distinct Teacher Professional Learning (PL) "series" for prekindergarten teachers on children's learning. The PL series offer teachers in a given site training on: 1) an evidence-based math curriculum (Clements & Sarama, 2008) and research-based interdisciplinary units developed by the NYC DOE; 2) integrating the arts (visual arts, music, dance, theater) into instruction; 3) integrating strategies drawn from an evidence-based program known as ParentCorps (Brotman et al., 2011) for supporting family engagement, child social-emotional development, and trauma-informed care; or 4) topics aligned to the district's quality standards (i.e., business as usual in this system). The team originally planned to leverage child-level lotteries that occur within NYC's deferred acceptance algorithm for preschool seats. They intended to identify lottery "winners" and "losers" for the three contrasts of interest (i.e., each PL series versus business as usual) and to collect data via direct assessments in preschool and kindergarten on approximately 800 lottery children per contrast (2400 children total across the three contrasts). However, they found a

number of methodological challenges with the child-level lottery design. Most importantly for the purposes of this paper, they found that site characteristics (such as quality and site type) were correlated with PL series (due in part to the fact that site assignment to PL is not entirely random; e.g., sites' PL preferences are taken into account when the NYC DOE assigns them to a PL series, in ways that would not allow them to isolate the effect of PL series from other sites characteristics). Subsequently, they pivoted to a cluster randomized design, with sites randomized into different PL series, which was a substantially stronger design for testing their research question about PL specifically.

IV. Challenges and Possible Solutions

As these five studies exemplify, there is considerable opportunity to leverage naturally occurring lotteries to answer pressing questions facing at-scale early education programs. However, there are a set of challenges that must be handled carefully in the design and analysis of these studies for their full potential to be realized. We discuss each challenge and possible solutions below, drawing on examples from the ongoing early education lottery studies described in the previous section.

IV.A. Challenge #1: Limited child-level covariates

Problem. Information on study participants' baseline characteristics is an essential part of studies that aim to identify causal impacts. In randomized designs, baseline covariates are used to assess *internal validity*, meaning whether the treatment and control groups are equivalent at baseline and for confirmatory outcomes at follow up (i.e., whether random assignment worked and whether differential attrition created a biased sample; Murnane & Willett, 2010). Covariates also can *increase statistical power* by explaining some of the residual variance in the relevant

outcomes. This can result in cost and time savings by reducing required sample sizes, as well as in more precise treatment effect estimates. Covariates also may be used to examine the *heterogeneity* of treatment effects (Bloom & Michalopoulos, 2013). For example, children from families with low incomes, dual language learners, and Latino children in particular tend to benefit more from public preschool than their peers (Phillips et al., 2017). Baseline measures of such key dimensions allow researchers to examine whether the effects of a given early childhood intervention similarly vary. Finally, covariates are critical for examining *external validity*, or to whom impact estimates apply – a topic we cover in more detail in Challenge #5 below.

Covariate information for lottery-based early education studies tends to be sparse. To illustrate this point, we display available covariate information for the five lottery-based studies in Table 2. As shown, arguably the richest and most useful covariates – student's prior test scores – are not collected at the time of preschool application in these (or to our knowledge, any) early childhood system that uses a lottery-based choice process. In some contexts, data are especially sparse due to efforts to reduce administrative burdens of application and to improve the equity of take up. For example, in DC, only students' age, address, and language of application are available for all applicants. In New York City, detailed demographic and screening data covariate information is available only on preschool applicants who subsequently enroll in preschool.

In contrast, lottery-based K-12 education studies tend to have much richer data available. For example, studies of New York City's Small Schools of Choice program had nine years of administrative data on applicants, covering basic student demographic characteristics, such as age, race, ethnicity, free-/reduced-price lunch status, English language learner status, and special education status, and scores from students' prior New York State standardized tests, such as 7th-

and 8th-grade English Language Arts and Mathematics (Bloom & Unterman, 2014). These data permitted that study team to illustrate empirically that random assignment "worked," providing two equivalent treatment and control group samples at baseline, as well as to examine whether balance was maintained throughout the follow-up period. In addition, that study team used these baseline data to compare students in the lottery sample with other students attending New York City Small Schools of Choice, as well as other high school students across the New York City School District. Furthermore, these data have enabled policy-relevant student subgroup analyses of variation in impacts, exploring for example, whether Small School of Choice impacts differed for students that entered high school performing below grade level in Mathematics and English Language Arts than for students who had previously performed at higher levels. Finally, these rich covariate data – especially highly predictive prior test scores – enabled the study team to conduct a rigorous propensity-score matching analysis and estimate the effects of Small School enrollment for all students attending SSCs, not just those who were in an SSC lottery, thereby helping to broaden the population of children who were studied.

Possible solutions. Avenues for addressing this issue include building collection of richer data into the preschool application process, adding baseline parent surveys, and adding pretests. Taking each in turn, research teams could work with a locality to add additional questions to application in-take forms. For example, a locality could ask parents to report on maternal education or family income when applying to its prekindergarten program (as in New Orleans). Of course, any additions must be balanced against administrative burden for participants and equity issues. Research has already shown that some of the groups most likely to benefit from public preschool programs are the least likely to apply (Shapiro, Martin, Weiland, & Unterman, 2019) and that administrative burden is a barrier for some families

interested in public preschool programs (Weixler, Valant, Bassok, Doromal, & Gerry, 2020). New Orleans has tried to strike this balance by emailing parents an optional survey after they submit their school choices as an additional data collection mechanism.

Collectively, we have found that public systems have not been able or willing to make changes to their application processes due to costs, logistics, privacy, and potential equity issues. Accordingly, some research teams have turned to baseline surveys for a subset of applicants to gather such data (see Table 2). Baseline surveys add cost to studies and are difficult to administer to all applicants. Parent surveys too, when not required for school entrance, typically have lower response rates and can be biased towards groups more likely to complete them. In addition, teams may have to wait until post-randomization to collect such data which is not ideal as randomization can influence families' responses and willingness to participate (Murnane & Willett, 2010). For example, the DC team is planning to administer a family survey to gather richer data. These surveys will be collected post-randomization as the team will need to know which families were randomized due to the intricacies of the City's assignment process writ large. They will administer the survey too to a subsample due to costs; thus they also need to know who was actually randomized to draw their subsample. One possible solution is that if localities help parents complete applications in centralized locations as Boston, DC, and New Orleans do, this process might be feasibly leveraged in future studies to consent parents and facilitate survey completion among all families or among a sample representative of the full range of program applicants.

The lack of child-level pretest data in these systems is important because child-level pretest data tends to explain more of the variation in child-level outcomes than other covariates,

providing more of a statistical power boost.[4] Pretest data also can provide more convincing evidence of baseline balance by treatment status and be used to create subgroups to test whether, as in prior literature, young children with lower pretest scores show larger gains then their peers in public preschool studies (Bitler, Hoynes, & Domina, 2014; Bloom & Weiland, 2015). Currently, three research teams are planning to collect these data prospectively in their lottery-based studies, using external trained data collectors (see Table 3). One team that is not – Boston – attempted prospective data collection in a lottery-based study before the pandemic. However, they faced power limitations due to high numbers of control crossovers that were compounded by the fact that due to small lottery blocks, non-consenting students resulted in incomplete blocks that could not contribute to estimates of treatment impacts. However, a large-scale study of Tulsa's Pre-K program was able to enlist teachers to assess all incoming students just before school began, at teacher meet-and-greet sessions with each individual child (Gormley et al., 2008). In lottery-based studies, it may be possible to similarly enlist school personnel for pretest assessments or to use the state-mandated direct assessments of children's school readiness in place in some states for this purpose. Due to logistical limitations, pretest assessments in such cases may have to occur after random assignment but before the intervention begins. This is not ideal timing since treatment assignment in theory may influence scores even before the intervention begins (Murnane & Willett, 2010). But such data would still be very valuable for the reasons we have outlined (enhancing internal validity, statistical power, and external validity plus making it possible to study the heterogeneity of impacts).

[4] As an example of the predictive power of pretests in ECE studies, in re-analysis of the Head Start Impact Study, Bloom and Weiland (2015) found that the pretest alone for explained 58% and 39% of the variance in vocabulary and early literacy outcomes, respectively, with a very marginal increase (7 and 3 percentage points, respectively) when then adding in the set of child- and family-level covariates used in the original study.

IV.B. Challenge #2: Limited data on the counterfactual

Problem. Multiple evaluation frameworks emphasize the importance of identifying not just whether an intervention "works" but whether it works compared to a well-identified counterfactual condition (Murnane & Willett, 2010; Weiss, Bloom, & Brock, 2014). Past empirical studies of early childhood programs provide rich illustrations of why this is important. For example, using a principal stratification framework, Feller and colleagues (2016) found that the effects of Head Start depended on what child care was like under the counterfactual condition, with effects concentrated in the subgroup of children who would have stayed home if they were not offered Head Start. In addition, Duncan and Magnuson (2013) demonstrate descriptively that since the early days of public preschool evaluation in the 1960s, immediate post-treatment impacts have declined and that the much greater availability of alternative programs is a prime explanation for why.

Identifying the treatment-control contrast is critical to all education studies. In ECE research, studies generally can identify what treatment group members experienced through available program and study-collected data. But identifying what control group children experienced is often more difficult in ECE than K-12 research. This is in part due to the U.S. policy context. For example, once children turn five in the U.S., they are eligible for free public education and the vast majority of these children enroll. Consequently, their educational settings are tracked by public data systems. In contrast, ECE is voluntary and supports for ECE data systems are fragmented and uneven across the country (Chaudry et al., 2021). The counterfactual accordingly tends to consist of a wider range of settings than in K-12 studies, with less administrative data to describe the mix of alternatives in a given context.

In some systems, families do in fact provide information on children's care settings at age 4 when they register for kindergarten (see Table 2). In Boston, these data were useful for understanding the alternative care settings for those children who competed in a lottery, lost the lottery, and ultimately did not enroll in the Boston program (i.e., the control compilers; Weiland et al., 2020). These data too allowed the team to identify the alternative care settings for all applicants who did not enroll, regardless of whether they participated in a lottery for an oversubscribed school. As shown in Figure 2, nearly all of the lottery control compilers in the Boston study enrolled in some out-of-home care, with nearly half in private settings and 88% in another preschool program. Among all applicants, the mix of settings was different, with fewer kids in other preschool programs and in different types of programs. These data were essential for interpreting the causal impacts of the Boston program and assessing their generalizability.

Ideally, to interpret study results, we would have information not just on alternative care setting type but about important features of the child-care setting like its quality, curriculum, and teacher qualifications. But here too, the U.S.'s decentralized, fragmented early education system means such data are rarely available. This issue is not unique to lottery-based early education studies; other study designs often face this challenge too. But this is another area where lottery-based early education studies are at a disadvantage versus studies of older children, in which many features of K-12 public schools are already centralized and publicly available.

Possible solutions. Data on counterfactual child care for studies of preschool programs can be gathered similarly to covariates – via building in questions in the registration process for kindergarten (as Boston and New Orleans do) and/or through surveys of families. The Boston example suggests that gathering both the name of the program and its type is beneficial for

cleaning and verification purposes, as is the use of pre-populated lists with validated names and types.

Data on the *features* of alternative care settings could be gathered via surveying these settings once they are reported by parents. If data are gathered early in the year before kindergarten, when children are enrolled in the alternative setting, observational quality assessments might also be possible. In the Head Start Impact Study, for example, the study team collected such data on the settings of control group children not enrolled in Head Start (Puma et al., 2012). These data were used in subsequent analyses to understand the contribution of treatment-control contrast in program quality to impacts on children (Friedman-Krauss, Connors, & Morris, 2016). Such data require substantial additional funding to gather but should be prioritized by funders and researchers in future lottery-based studies of early childhood when possible.

IV.C. Challenge #3: Limited outcome data

Problem. To our knowledge, all published lottery-based studies have leveraged administrative data to obtain outcome measures for their samples of interest. For example, a study of the impacts of Michigan's largest charter school network used state records of students' math and reading test scores, grade retention, special education placement, and disciplinary incidents in grades 3-8 (Dynarski et al., 2018). Another charter school study leveraged participants' voting records to explore the effect of education on civic participation (Cohodes & Feigenbaum, 2021). Other such prominent examples include New York City Small Schools of Choice (Unterman & Haider, 2019), which leverages district administrative records for grades 9-12, National Student Clearinghouse Data for postsecondary enrollment records and degree

attainment, and New York State Unemployment Insurance Data for employment and earnings outcomes.

However, sometimes, there are gaps in *what* is available for outcome measures and *when* it is available for lottery-based studies. For example, the Michigan charter study did not include measures of children's moral character, a central focus of the charter network (Dynarski et al., 2018). In addition, there were lotteries in that study that began in kindergarten but some outcome measures – like math and reading test scores – were not available until third grade. Consequently, that research team could not identify whether there were different or cumulative effects across grades for children in the early grades.

These issues of *when* and *what* are particularly pronounced in all early education studies that rely on administrative data, not just lottery-based studies. For example, in propensity-score-based studies of Tulsa's Pre-K program that rely on state and district records and difference-in-difference studies of state pre-k programs that use the NAEP, academic outcomes are not available until third or fourth grade (Fitzpatrick, 2008; Hill, Gormley, & Adelstein, 2015). This timing is problematic given considerable evidence that the largest benefits of a given preschool program occur at the end of the program and may no longer be detectable by the end of kindergarten on widely used measures in the field (Lipsey et al., 2018; Puma et al., 2012). Evidence also shows that whether the preschool boost is sustained can depend on children's educational experiences in the early elementary years (Johnson & Jackson, 2019; Mattera et al., 2022; Unterman & Weiland, 2020). But without data on children's outcomes before third grade, we cannot discern between programs with no impact at all from programs with a strong initial impact that faded due to subsequent experiences. The practice and policy implications in the two

scenarios are very different, making this limitation a major one for evidence-based improvement efforts.

On the *what* (or substance) side, the best evaluations are theory-based (Murnane & Willett, 2010). In early education, they engage deeply with theoretical frameworks on *how* early education programs support children and families, in which domains, and through which contextual mediators and moderators. Educational administrative data generally lack measures of possible mediators and moderators, as well as some of the key outcomes of early education programs such as child social-emotional development, behavior, family engagement and maternal employment. Accordingly, studies that rely on administrative data only available in public education systems may miss or underestimate the potential benefits (or not) of these programs.

Possible solutions. Recognizing the limitations of the timing and content (i., *when* and *what)* of outcomes available in administrative data, some of our five teams have begun or are planning *prospective* data collection with direct assessments of young children. As shown in Table 3, for example, the Montessori team is collecting outcome data on children at the end of children's three- and four- year old preschool years and at the end of kindergarten. Their work includes widely used measures in the field of children's math, language, and early literacy that will permit cross-study comparability. They are also collecting more novel data on children's skills that match the unique theory of the Montessori model – i.e., persistence and a mastery orientation. The DC team too plans to collect widely used measures of children's language, literacy, math, executive functioning, and social emotional skills to compare results to other early childhood impact studies. They also plan to add measures of children's racial attitudes new to preschool evaluation, following one of the hypothesized benefits of DC programs. That is,

because child care and early education programs are more segregated than K-12 settings (Greenberg & Monarrez, 2019), the study team hypothesizes that school-based preschool – universally available and administered by lottery – may be more racially mixed than available alternatives and have the institutional support necessary to address early explicit bias. To our knowledge, these dynamics have not yet been studied. However, research shows that children can distinguish between racial groups by three months, show favorable attitudes toward their own racial group by nine months, and employ racial stereotypes by six years, making public preschool a potentially important time to support the development of inclusive social skills and intergroup attitudes (Kelly et al., 2005; Lee, Quinn, & Pascalis, 2017; Pauker, Ambady, & Apfelbaum, 2010).

Notably, however, prospective outcome data collection can be very difficult in lottery-based early education studies. The Boston Alignment team, for example, ultimately decided against attempting prospective data collection via direct child assessments. Preschool blocks can be quite small compared to those in K-12; losing just a few families across blocks can result in incomplete blocks and then worsen both statistical power and external validity issues. Differential attrition in particular was too large of a risk, given that families who lost the lottery were not particularly motivated to participate in assessments. Consent rates too might have varied substantially across blocks, presenting design decisions around who to sample and include. In addition, if compliance is relatively low, very large numbers of participants are needed to generate sufficient statistical power to detect intervention effects.

Enriched administrative data may be another possibility. Many school districts are now adopting benchmark assessments to monitor student progress in the early grades. Some state laws even require such assessments, such as third-grade reading laws. For example, the

Michigan Education Data Center (MEDC) is gathering and cleaning such data from benchmark assessments required by the state's Third Grade Reading law. Unlike third grade and up state standardized tests, districts tend to have leeway in which benchmark or progress monitoring assessments they choose, which can lead to inconsistent outcomes available for preschool studies. For example, Boston used an early reading assessment called DIBELS for many years while surrounding districts did not and further, such data were not compiled at the state level. A study of Boston Prekindergarten that leveraged these data could do so accordingly only for children who remained in BPS schools (Weiland, Unterman, & Shapiro, 2021). But when available and when equivalent across districts, these data offer promise for providing more timely, policy relevant evidence on the effects of preschool programs.

IV.D. Challenge #4. Attrition

Problem. As mentioned previously, empirical studies have shown that the lotteries generated by these school choice systems have strong *internal validity* – i.e., they result in treatment and control groups at baseline who were essentially randomized in a coin-flip-like procedure and who are equal in expectation before a given intervention began (e.g., Bloom & Unterman, 2014; Gray-Lobe et al., 2023). However, a more vexing problem – as it tends to be for most studies in education that in principle, can identify causal effects – is attrition (i.e., when students disappear from the follow-up dataset). That is, to be fully credible, researchers must show: 1) that there has not been differential attrition by treatment status; and 2) that there is still balance in baseline characteristics for the non-attritors (Krueger & Zhu, 2004; Murnane & Willett, 2010). Both analyses are easy to conduct analytically and are standard in empirical research. But when evidence of biasing attrition is found, there are no simple fixes that can fully restore confidence in the internal validity of a study's impact estimates.

Issues of attrition can be exacerbated in lottery-based early education studies for several reasons. First, features of systems play an important role. In some preschool systems, like in New York City, students are only given a unique identifier that follows them through 12th grade if they enroll in public preschool. Students who apply but do not enroll can receive a unique identifier if they enroll later, in kindergarten or beyond. But matching them to their preschool enrollment records requires additional matching processes that are resource-intensive. In New York City, about 11-18% of prekindergarten applicants who participated in a lottery for an oversubscribed site did not enroll in any prekindergarten slot. There was evidence this occurred differentially, with 11-16% of lottery winners not enrolling versus 16-18% of lottery losers. Unfortunately, in this instance, there is a differential attrition issue, but no demographic data available on the children that are missing outcome data, making it very difficult to assess the extent of the attrition-induced bias. In contrast, in an instance like the study of New York City's Small (High) Schools of Choice, when students choose to leave the district after participating in a lottery, their demographic and prior academic achievement data is available and extensive sensitivity tests are possible (Bloom & Unterman, 2014).

Second, the early childhood years are when families are often more mobile than when their children are older. Accordingly, in many contexts, families of young children may be more likely to move out of a given locality, particularly if they do not receive a school they would like their child to attend via a lottery system. If statewide data are available, children can be followed into other localities (via either a unique identifier or an additional matching process otherwise). But if not, differential attrition can be a difficult problem. For example, preliminary evidence shows that about 69% of children who applied to DC's preschool program for three year olds and participated in a lottery were enrolled in DC public schools in kindergarten two years later. The

31% who were not are lost to the study team using in DC administrative data. As we show in Appendix Table 1, there was evidence of differential attrition by treatment status, though this difference is relatively small (about 4 percentage points) when controlling for the likelihood of being matched to a three year old program. In Appendix Table 2, early evidence shows that balance was fairly well maintained on the limited baseline characteristics available.

Possible solutions. Common advice in the education research field is try to avoid attrition and when you cannot, do your best to understand it (i.e., who attritted and why; Murnane & Willett, 2010). On the prevention side, research teams facing differential attrition problems can work to create robust longitudinal datasets that span multiple school districts and states. In addition, researchers can also encourage states and localities to assign a unique identifier at preschool application (or even birth) to allow for more seamless tracking of children for research purposes.[5] Finally, on the understanding attrition side, collecting richer baseline data on student demographics and pretests as we discussed in Challenge #1 can allow for deeper insight into which students are attriting and thus better assessment of the potential effects of attrition on internal validity.

IV.E. Challenge #5. External validity

Problem. All empirical education studies have to contend with external validity, or to whom impact estimates apply. If effects are heterogeneous, results of a given study generalize only to the population they represent (Murnane & Willett, 2010). For example, if a research

[5] There is also recent methodological work that may help study teams that lose large portions of their sample move forward with the data that they have. For example, Weidmann and Miratrix (2021) used data from 10 randomized controlled trials to assess the magnitude of the bias that occurs when varying amounts of follow-up data are available. They found that when attrition occurs equally between groups, the bias is smaller than originally anticipated.

team randomly sampled students from only elementary schools in the northern end of a district, the subsequent study's results apply technically only to elementary school students in elementary schools in that part of the district. They do not apply to middle school students in that same district, to elementary school students in another district, not to elementary school students in other schools in the same district. The reason is that students in the study may differ from other students in ways that make an intervention, program, or policy affect students in that district differently than students elsewhere (i.e., effects may be heterogeneous). In empirical research, determining to whom the researcher would like to generalize is a critical step in making sampling decisions.

Methods for assessing external validity are generally quite simple.[6] Researchers compare the characteristics of participants and settings on average in their study to those of the population. Similar characteristics indicate that study results are more applicable to the population; differences indicate that results are less generalizable.

In lottery-based early education studies, the core external validity issue is that the lotteries are *naturally occurring*, within oversubscribed programs. Researchers have no control over who is ultimately randomized; external validity is not a study design feature that can be manipulated by the research team to answer the question of interest. Rather, after randomization occurs, the research team then learns who was randomized and thus to whom studies that leverage this randomization would apply. Entire schools (and the students who applied to those schools) can be left out of a given sample as well, if they were not over-subscribed.

[6] Newer work has shown other, more intensive approaches to assessing generalizability (Tipton & Olsen, 2018; 2022; Stuart, Ackerman, & Westreich, 2018). These methods offer other ways to parameterize the problems we describe but do not solve them. Accordingly, we stick to simpler methods in our discussion here.

So far, external validity findings from preschool lottery studies show that this issue can have major implications for study design and interpretation. For example, in Washington DC, from 2014-2018, around 25,197 families applied for a three-year-old seat and 5,997 ultimately competed in a lottery (24%). As shown in Table 4, there were large differences in neighborhood income, racial composition, and educational attainment when comparing all applicants, the randomized sample, and those who complied with their lottery assignment. For example, median neighborhood income for applicants was about $81K versus $107K for the randomized sample and $141K for compilers. Nearly half of lottery compilers are drawn from just one ward or neighborhood (Ward 6), even though only 16% of applicants live in this ward.

External validity findings from the study of Boston's rollout of an aligned prekindergarten and kindergarten curriculum offer interesting evidence that suggest that the design may address some questions better than others (McCormick et al., 2022). An earlier lottery-based study of the effects of Boston Prekindergarten versus alternatives (Weiland et al., 2020) found substantial differences in background characteristics between those randomized in the lottery process versus the full set of applicants. For example, among randomized applicants, 51% qualified for free-reduced-priced lunch and 28% were White, versus 65% and 17% of all applicants, respectively. Lotteries were also highly concentrated in a subset of schools. Accordingly, the authors took care to caveat that their study results applied to more advantaged students who wanted to attend a subset of oversubscribed district schools and not effects for the full set of students who wanted to attend. In contrast, as shown in Appendix A Table 3, applicants to schools implementing aligned curriculum and applicants who participated in a lottery were much more similar to the full set of applicants (e.g., 67% of applicants were eligible

for free-reduced-priced lunch, versus 68% for applicants to aligned schools and 61% of the lottery sample).

Possible solutions. Given researchers' lack of control of the randomization process, lottery-based studies may answer a different question than the research team intended at the outset – a problem of which limited external validity is a symptom. A way forward in improving external validity with lottery-based early education studies is to obtain prior lottery data and covariates information to first understand in past years who was randomized in a given system and what settings are represented in the randomized subset of applicants (assuming similar processes from one year to the next). The design of future such studies should be informed heavily by these analyses, so that researchers can be more certain of what research questions they can address with data from these systems and determine whether these are the policy questions of interest. This will likely require more funder support for less definitive, exploratory analyses.

These early-stage analyses can also help build the case for alternative designs. Lottery-based designs are attractive because they do not disrupt localities' normal operations. However, presenting information that data from these systems may not answer the question of interest may help to persuade decision makers to allow other designs, like randomizing classrooms or schools, that can better answer the research questions of interest.

If previous data show that lotteries from these systems can answer questions of interest for a locality and the broader field, external validity can be assessed following models from K-12 and from preschool specifically. For example, Abdulkadiroğlu and colleagues (2011) provide an excellent road map in their lottery-based study of charter schools for assessing to whom study results are likely to generalize, including a lottery-based propensity score validation approach for examining whether students not in the subsample randomized would likely experience benefits if

enrolled in charter schools instead of alternatives. The first lottery-based preschool study (Weiland et al., 2020) followed and extended this example, ultimately examining the concentration of lotteries in certain schools, the characteristics of the lottery subsample versus all applicants, differences in the counterfactual between the lottery subsample control group versus all applicant non-enrollees, and whether lottery impact findings likely generalized to all applicants (they did not).

External validity work does depend on having good covariate, counterfactual, and education setting data at hand. Our possible solutions to those challenges also apply for addressing and solving external validity challenges.

IV.F. Challenge #6. Answering site-level questions with child-level randomization

Problem. As public preschool programs have become more common, there has been increasing interest in not just whether to fund preschool but how to make it more effective (Weiland, 2018). Localities that administer these programs tend to be particularly interested in such questions. Should they hire teachers with BA? Should they continue using their current curriculum or switch to an alternative? What is the best assessment system for providing actionable, feasible, and valid information on student learning?

Teams are just beginning to explore when and how to leverage the preschool lotteries created in school choice systems to address such site-level questions. As described, one of our teams is using student-level lotteries to examine the impact of Boston's rollout of an aligned prekindergarten and kindergarten curriculum on students' learning in third grade. And in New York, the City was interested in estimating the impacts of different professional learning (PLD)

for preschool teachers on student learning in its universal preschool system. NYU researchers initially proposed using child-level lotteries to do so.

Ultimately, the New York City team found that they could not answer the City's questions using the child-level lotteries from the deferred acceptance system. Sites had selected into PLD series and one of several methodological challenges was that series were associated with other characteristics of sites.[7] This is not surprising in a system that relies on program leaders' rank-ordering preferences among the PL series, as well as site need and series capacity, to make PL series assignments.

But estimates leveraging the child-level lotteries accordingly would represent the *joint impact* of all the characteristics of sites, not just their PLD series. Said differently, it would be impossible to disentangle the effect of each PLD series track from site type, the children that attend these sites, and the assessed CLASS quality scores of the site prior to participating in the series. The study team learned that in fact, site-level randomization had occurred due to constraints on capacity for each series. They pivoted to leverage this source of randomization and to conduct a cluster randomized trial instead.

The Boston team grappled with similar site-selection issues as schools selected into implementing the aligned curriculum (or not) per the district's autonomous schools model. Site characteristics were similarly correlated with alignment status. In designing their study, the team accordingly was careful to be clear they were testing not the effects of alignment on its own but the district's *rollout of an aligned curriculum.* Given the paucity of causal evidence on this

[7] Other methodological concerns included differential attrition from preschool, as discussed earlier; challenges with "non-compliance" as lottery losers enrolled in sites in the series for which they served as controls; a "mixed counterfactual" as each treatment series would be compared to a number of other series, making interpretation difficult; and limited power.

topic, the district and research team felt the study would still answer a vital question, even if it could not isolate the effects of alignment alone. This issue is akin to other circumstances in which a set of schools are targeted for additional resources due to low student achievement levels. For example, in studies of the effects of School Improvement Grant (SIG) funds for schools with chronically low academic achievement, researchers used various analytic approaches to estimate the effects of SIG funds, while acknowledging that any effects may also be the result of a package of supports that schools attract when in need of intervention (Dragoset et al., 2017, LiCalsi et al., 2015, Dee, 2012).

Possible solution. This issue is critical to address in the design phase. Just as in our external validity solutions, a concrete way forward is to obtain prior system data to first understand in past years who was randomized in a given system and what settings are represented in the randomized subset of applicants. These data, along with close communications and interviews with staff in a given locality, can help to pinpoint *where* a setting-level intervention is implemented, the selection process into implementation, and site characteristics that may be correlated with a given intervention. These data and analytics can help the study team and locality understand what question the design can versus cannot answer. From there, a pivot may be in order (as in New York City) to a different research design.

V. Summary: Recommendations for designing preschool lottery studies

Our joint work on leveraging naturally occurring early education lotteries illuminates both the promise and challenges of this design in this new context. As we highlighted in our introduction, many of the challenges of this design are the same as in any empirical education study, particularly those aiming to identify causal relations. But some of these challenges are

exacerbated in lottery-based early education studies and require careful handling in study design, analysis, and interpretation.

For future such studies, we offer the following recommendations:

1) When designing lottery-based studies, start with the program's theory of change, a locality's research questions, and gaps in the broader research evidence base. The highest quality and most useful educational empirical studies for guiding policy and practice tend to combine these three essential elements when identifying research questions.
2) Identify the covariates, outcomes, and counterfactual data that are available from administrative data. Use field-based efforts and supplements to the preschool application process, to address any important gaps in this data, such as the lack of rich covariates and lack of a measure of the key outcome that the program was supposed to move.
3) To limit attrition problems, consider opportunities to create robust longitudinal datasets that span multiple school districts and states and set up systems for tracking students across localities using a common identifier from the time of preschool application.
4) Anticipate the external validity of a lottery-based study from past years' data and use it to determine a priori what research questions a lottery-based study can answer well and which ones require a different design. Because the pandemic has changed enrollment patterns for young children especially (Bassok & Shapiro, 2021; Greenberg, 2021; Weiland et al., 2021), studies with cohorts *after* the pandemic began might be better informed by data from cohorts from 2021 onwards than by data from pre-pandemic cohorts.

5) There are tradeoffs to consider in choosing an analytic strategy for estimating impacts from a lottery-based design. In particular, one must choose samples drawn from first-choice lotteries for children, first lotteries for children, and assignment score approaches. For example, the assignment score approach in theory may improve external validity but it may not permit predicting cross-site variation since two students with the assignment score (block) may have applied to different schools with different characteristics (Bloom et al., 2017). More research comparing these approaches directly in the preschool space is needed. Teams should weigh the tradeoffs between them carefully, determine which best answers their particular research questions, and, as a robustness check, ideally conduct the analysis multiple ways.

6) For site-level questions, pinpoint *where* a setting-level intervention is implemented, the selection process into implementation, and site characteristics that may be correlated with a given intervention. This work is critical as site characteristics can be confounded with the main characteristic of interest. Pivot to a different research design, if child-level randomization cannot satisfactorily answer a site-level question.

7) Find opportunities to connect with colleagues engaged in similar work. Collaboration between our five teams began organically, with researchers considering a lottery-based design connecting with those who were already in the process of doing so. A conference grant from the Spencer Foundation provided us with resources to more formally engage with one another. As teams leverage lotteries in other contexts and to address other questions, similar collaborative networks have a role to play in improving applied studies and accordingly shaping future evidence-based policy and practice.

8) Finally, we also hope that funders will begin to recognize the potential contributions of the lottery-based design for building the next generation of evidence on early education programs. Funding for the early stage work to identify what questions these lottery-based early education studies can answer in a given context and the relevance of those questions to practice partners is essential. Prospective field work too in these studies can be very challenging and may require additional resources, beyond those required in other kinds of studies that can identify causal impacts. We hope illuminating the particularities and nuances of the design across our five studies can also inform funder priorities and decisions.

Rigorous design has long characterized early education studies, dating back to the landmark Perry and Abecedarian studies in the 1960s and 1970s. And since around 2000, there has been a dramatic rise in the use of methods that can identify causal effects of education programs, practices, and policies more broadly. In addition to improving early education studies directly, we hope that our joint work also serves as a case study of how educational context can affect study design when moving a study design into a new educational topic area.

References

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *The Quarterly Journal of Economics*, *126*(2), 699-748.

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., & Pathak, P. A. (2017). Research design meets market design: Using centralized assignment for impact evaluation. *Econometrica*, *85*(5), 1373-1432.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton University Press.

Bassok, D., & Engel, M. (2019). Early childhood education at scale: Lessons from research for policy and practice. *AERA Open*, *5*(1).

Bassok, D. & Shapiro, A. (2021). *Understanding COVID-19-ear enrollment drops among early-grade public schools students.* Brookings Institute. https://www.brookings.edu/blog/brown-center-chalkboard/2021/02/22/understanding-covid-19-era-enrollment-drops-among-early-grade-public-school-students/

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start* (No. w20434). Cambridge, MA: National Bureau of Economic Research.

Bloom, H. S. (Ed.). (2005). *Learning more from social experiments: Evolving analytic approaches*. Russell Sage Foundation.

Bloom, H. S., & Michalopoulos, C. (2013). When is the story in the subgroups? *Prevention Science*, *14*(2), 179-188.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness, 10*(4), 817-842.

Bloom, H. S., & Unterman, R. (2014). Can small high schools of choice improve educational prospects for disadvantaged students? *Journal of Policy Analysis and Management, 33*(2), 290-319.

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*. New York, New York: MDRC.

Brotman, L. M., Calzada, E., Huang, K. Y., Kingston, S., Dawson-McClure, S., Kamboukos, D., ... & Petkova, E. (2011). Promoting effective parenting practices and preventing child behavior problems in school among ethnically diverse families from underserved, urban communities. *Child Development, 82*(1), 258-276.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal, 17*(2), 372-404.

Chaudry, A., Morrissey, T., Weiland, C., & Yoshikawa, H. (2021). *Cradle to Kindergarten: A new plan to combat inequality*. New York, NY: Russell Sage.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal, 45*, 443-494.

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.

Cohodes, S., & Feigenbaum, J. J. (2021). *Why does education increase voting? Evidence from Boston's charter schools* (No. w29308). Cambridge, MA: National Bureau of Economic Research.

Dee, T. (2012). *School turnarounds: Evidence from the 2009 stimulus (No. w17990)*. Cambridge, MA: National Bureau of Economic Research.

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., ... & Giffin, J. (2017). *School Improvement Grants: Implementation and Effectiveness. NCEE 2017-4013*. National Center for Education Evaluation and Regional Assistance.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895-3962.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*(2), 109-32.

Dynarski, S., Hubbard, D., Jacob, B., & Robles, S. (2018). *Estimating the effects of a lage for-profit charter school operator* (No. w24428). Cambridge, MA: National Bureau of Economic Research.

Feller, A., Grindal, T., Miratrix, L., & Page, L. C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics*, *10*(3), 1245-1285.

Fitzpatrick, M. D. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The BE Journal of Economic Analysis & Policy*, *8*(1).

Friedman-Krauss, A. H., Connors, M. C., & Morris, P. A. (2017). Unpacking the treatment contrast in the Head Start Impact Study: To what extent does assignment to treatment affect quality of care? *Journal of Research on Educational Effectiveness*, *10*(1), 68-95.

Gormley Jr, W. T., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science*.

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

Greenberg, E. (2021). *Better data use shows the depths of the pandemic prekindergarten crisis.* Washington, DC: Urban Institute. https://www.urban.org/urban-wire/better-data-use-shows-depths-pandemic-prekindergarten-crisis

Greenberg, E., & Monarrez, T. (2019). *Segregated from the start: Comparing segregation in early childhood and K–12 education*. Urban Institute, https://www. urban. org/features/segregated-start.

Harding, J.F., McCoy, D., & McCormick, M. (2020). Understanding alignment in children's early learning experiences: Policies and practices from across the United States. *Early Childhood Research Quarterly*, *52,* 1-4.

Hill, C. J., Gormley Jr, W. T., & Adelstein, S. (2015). Do the short-term effects of a high-quality preschool program persist? *Early Childhood Research Quarterly*, *32*, 60-79.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy, 11*(4), 310-49. Doi: 10.1080/19345747.2018.1441347

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Gibson, A., Smith, M., ... & Pascalis, O. (2005). Three-month- olds, but not newborns, prefer their own-race faces. *Developmental Science, 8*(6), F31-F36.

Krueger, A. B., & Zhu, P. (2004). Another look at the New York City school voucher experiment. *American Behavioral Scientist, 47*(5), 658-698.

Lee, K., Quinn, P. C., & Pascalis, O. (2017). Face race processing and racial bias in early development: A perceptual-social linkage. *Current Directions in Psychological Science, 26*(3), 256-262.

LiCalsi, C., Citkowicz, M., Friedman, L. B., & Brown, M. (2015). *Evaluation of Massachusetts Office of District and School Turnaround Assistance to Commissioner's Districts and Schools: Impact of School Redesign Grants*. American Institutes for Research.

Lincove, J. A., Valant, J., & Cowen, J. M. (2018). You can't always get what you want: Capacity constraints in a choice-based school system. *Economics of Education Review, 67,* 94-109.

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.

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.

Love, J. M., Kisker, E. E., Ross, C., Raikes, H., Constantine, J., Boller, K., ... & Vogel, C. (2005). The effectiveness of Early Head Start for 3-year-old children and their parents: Lessons for policy and programs. *Developmental Psychology*, *41*(6), 885.

Mattera, S.K., Jacob, R., MacDowell, C., & Morris, P. (2022). *Long-term effects of enhanced early childhood math instruction: The impacts of Making Pre-K Count and High 5s on third grade outcomes.* New York, NY: MDRC.

McCormick, M., Unterman, R., Pralica, M., Weiland, C., Weissman, A., & Hsueh, J. (2022). *A new approach to sustaining pre-K impacts: Leveraging naturally occurring lotteries to examine a district-wide rollout of instructional alignment across pre-K and kindergarten*. New York, NY: MDRC. https://files.eric.ed.gov/fulltext/ED619526.pdf

Monarrez, T., Greenberg, E., Luetmer, G., & Chien, C. (2020). *Using centralized lotteries to measure preschool impact: Insights from the DC Prekindergarten Study*. Washington, DC: Urban Institute. https://www.urban.org/research/publication/using-centralized-lotteries-measure-preschool-impact

Morris, P., Mattera, S. K., Castells, N., Bangser, M., Bierman, K., & Raver, C. C. (2014). *Impact findings from the Head Start CARES demonstration: National evaluation of three approaches to improving preschoolers' social and emotional competence*. New York, NY: MDRC.

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

My School DC. (2019). *How My School DC matches students to schools.* https://www.youtube.com/watch?v=pBRiCyjiZao

National Center for Montessori in the Public Sector. (n.d.). *About Montessori.* Retrieved from https://www.public-montessori.org/montessori/.

Pauker, K., Ambady, N., & Apfelbaum, E. P. (2010). Race salience and essentialist thinking in racial stereotype development. *Child Development, 81*(6), 1799-1813.

Phillips, D., Lipsey, M., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., Duncan, G. J., Dynarski, M., Magnuson, K. A., & Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects.* Washington, DC: Brookings Institution.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. http://dx.doi.org/10.3102/ 0013189X0933237

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins,F., Mashburn, A., & Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study Final Report*, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Roth, A. E. (2008). Deferred acceptance algorithms: History, theory, practice, and open questions. *International Journal of Game Theory*, *36*(3), 537-569.

Sabol, T. J., McCoy, D., Gonzalez, K., Miratrix, L., Hedges, L., Spybrook, J. K., & Weiland, C. (2022). Exploring treatment impact heterogeneity across sites: Challenges and opportunities for early childhood researchers. *Early Childhood Research Quarterly*, *58*, 14-26.

Shapiro, A., Martin, E., Weiland, C., & Unterman, R. (2019). If you offer it, will they come? Patterns of application and enrollment behavior in a universal prekindergarten context. *AERA Open*, *5*(2).

Stein, A., & Coburn, C. E. (2021). Instructional policy from Pre-K to third grade: The challenges of fostering alignment and continuity in two school districts. *Educational Policy*.

Stuart, E. A., Ackerman, B., & Westreich, D. (2018). Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on Social Work Practice*, *28*(5), 532-537.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516-524.

Tipton, E., & Olsen, R. B. (2022). *Enhancing the Generalizability of Impact Studies in Education. (NCEE 2022-003).* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee.

Unterman, R. (2017). *An early look at the effects of Success Academy Charter Schools*. New York, NY: MDRC.

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

Unterman, R., & Haider, Z. (2019). *New York City's Small Schools of Choice: A first look at effects on postsecondary persistence and labor market outcomes*. New York, NY: MDRC.

Unterman, R., & Weiland, C. (2020). *Higher-quality elementary schools sustain the prekindergarten boost: Evidence from an exploration of variation in the Boston Prekindergarten program's impacts.* Providence, RI*:* Annenberg Institute for School Reform at Brown University, EdWorkingPaper No. 20-321.

Weidmann, B., & Miratrix, L. (2021). Missing, presumed different: Quantifying the risk of attrition bias in education evaluations. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(2), 732-760.

Weiland, C. (2018). Pivoting to the "how": Moving preschool policy, practice, and research forward. *Early Childhood Research Quarterly*, *45*, 188-192.

Weiland, C., Greenberg, E., Bassok, D., Markowitz, A., Guerrero Rosada, P., Luetmer, G., Abenavoli, R., Gomez, C., Johnson, A., Jones-Harden, B., Maier, M., McCormick, M., Morris, P., Nores, M., Phillips, D., & Snow, C. (2021). *Historic crisis, historic opportunity: Using evidence to mitigate the effects of the COVID-19 crisis on young children and early care and education programs.* University of Michigan Education Policy Initiative and Urban Institute Policy Brief. Retrieved from https://edpolicy.umich.edu/files/EPI-UI-Covid%20Synthesis%20Brief%20June%202021.pdf

Weiland, C., Unterman, R., & Shapiro, A. (2021). The kindergarten hotspot: Literacy skill convergence between boston prekindergarten enrollees and nonenrollees. *Child Development*, *92*(2), 600-608.

Weiland, C., Unterman, R., Shapiro, A., Staszak, S., Rochester, S., & Martin, E. (2020). The effects of enrolling in oversubscribed prekindergarten programs through third grade. *Child Development*, *91*(5), 1401-1422.

Weiland, C., Unterman, R., Shapiro, A., & Yoshikawa, H. (2019). *Findings on Boston Prekindergarten through early elementary school*. Ann Arbor, MI: Ford Education Policy Initiative Policy Brief. http://edpolicy.umich.edu/files/boston-prekindergarten-findings.pdf

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.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778-808.

Weixler, L., Valant, J., Bassok, D., Doromal, J. B., & Gerry, A. (2020). Helping parents navigate the early childhood education enrollment process: Experimental evidence from New Orleans. *Educational Evaluation and Policy Analysis*, *42*(3), 307-330.

White House. (2013, February 13). *Fact sheet: President Obama's plan for early education for all Americans.* Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2013/02/13/fact-sheet-president-obama-s-plan-early-education-all-americans

White House. (2021, April 28). *Fact sheet: The American Families Plan.* Retrieved from https://www.whitehouse.gov/briefing-room/statements-releases/2021/04/28/fact-sheet-the-american-families-plan/

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.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M., Espinosa, L., Gormley, W., Ludwig, J.O., Magnuson, K.A., Phillips, D.A., & Zaslow, M.J. (2013). *Investing in our future: The evidence base on preschool education.* New York: Foundation for Child Development and Ann Arbor, MI: Society for Research in Child Development.

Tables and Figures

Figure 1: School choice process for a hypothetical preschool applicant in a DA choice system

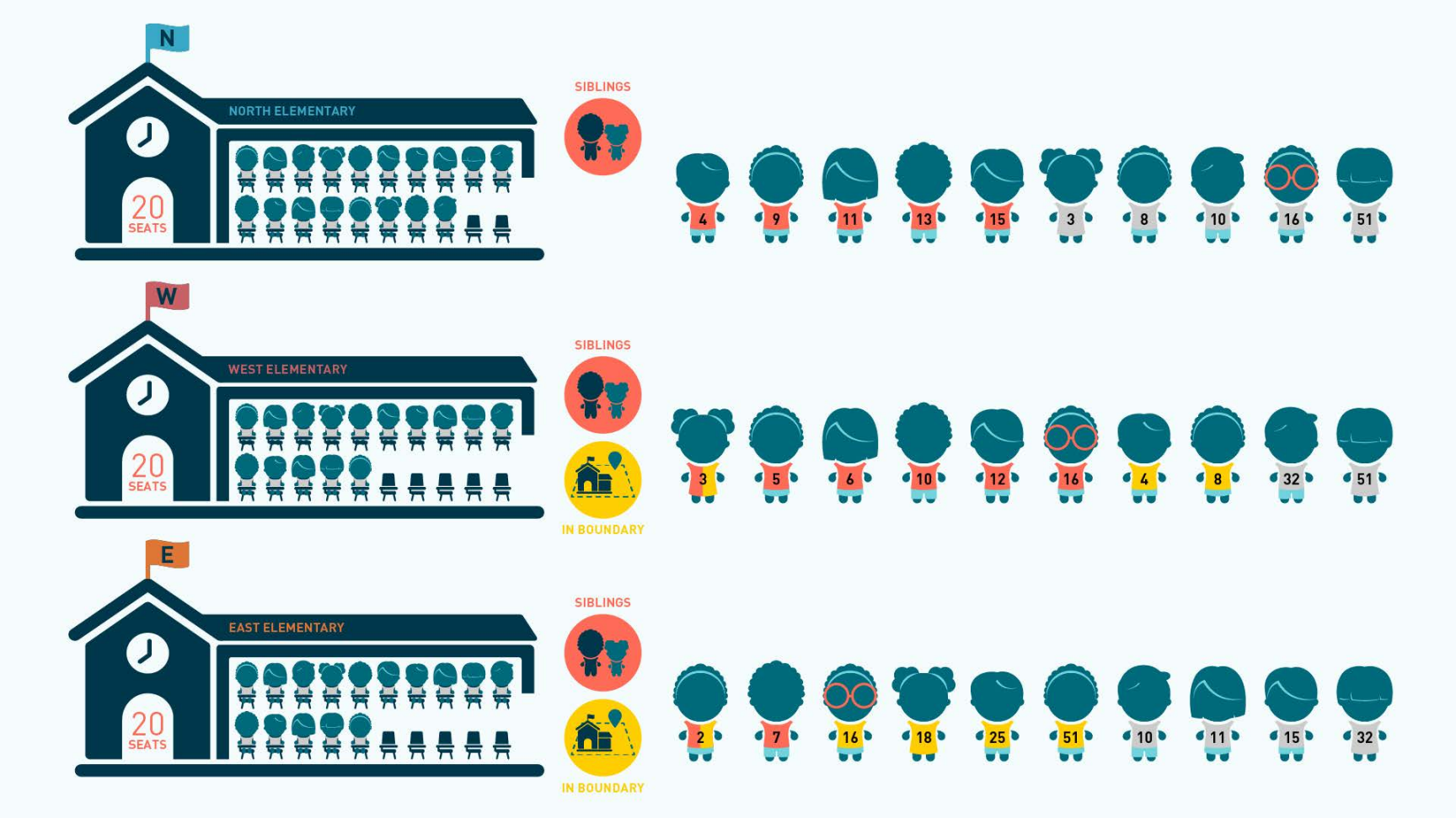
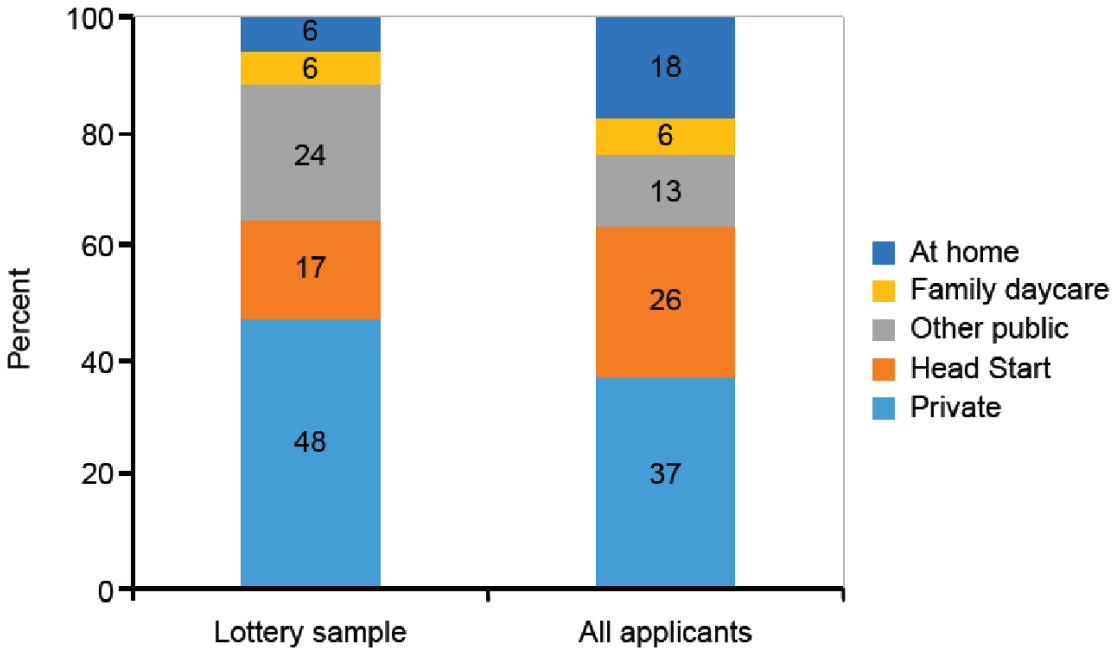


Figure 2: Non-Boston Prekindergarten care settings in the year before kindergarten for lottery sample control compliers versus all applicants



Source: Weiland, Unterman, Shapiro, & Yoshikawa (2019).

Table 1: Summary and key features of the five lottery-based preschool studies

Study details	**Boston**	**DC**	**Montessori**	**New Orleans**	**New York City**
Main research question	Effects of the district's rollout of prekindergarten and kindergarten curriculum alignment	Effects of DC Prekindergarten on 3 year olds	Effects of high-fidelity public Montessori for three-year-olds	1.Effects of higher-quality vs. lower-quality programs (as measured by CLASS; 2. Parents' first-choice PK vs. lower-ranked choices	Effects of three distinct professional learning (PL) "series" versus business-as-usual on children's school readiness
Lottery study cohorts	3 cohorts (2012-2014)	4 cohorts with admin data only (2014-2019); 2 cohorts with researcher-collected and admin data (2022 & 2023)	1 cohort (2022)	7 cohorts (2017-2023)	1 cohort (2016); 2 cohorts (2019 & 2020) planned but canceled due to pilot results
Treatment condition	Winning a lottery for an aligned school and enrolling in an aligned school	Winning a lottery for a 3 year old slot and enrolling	Winning a lottery for a high-fidelity Montessori program (and possibly, of enrolling)	1. Winning a lottery for the family's first-choice program 2. Winning a lottery for a highly-rated program	Three treatment conditions: Explore (math-focused teacher PL), Thrive (family engagement / social-emotional focused PL), and Create (arts-focused PL)
Control condition	Losing lottery for an aligned school and subsequently: - enrolling in another aligned school during a subsequent lottery round (i.e., cross-overs) - enrolling in an unaligned school (i.e., control compliers) - or not enrolling in BPS prekindergarten at all (i.e., also control compliers)	Losing lotteries for ranked schools and not enrolling in DCPS at age 3	Losing the initial lottery (ITT) and not enrolling in a high-fidelity Montessori program at age 3 (LATE)	1. Losing the lottery for the family's first-choice program 2. Losing lotteries for all highly-rated programs	Losing first lottery and: -enrolling in a program with a different PL series; (i.e. cross-overs) -enrolling in a program with the business-as-usual PL series (i.e., control compliers) -not enrolling in NYC Pre-K (i.e., control compliers but lost to follow-up)
Total N of applicants	10,318	4 cohorts with admin data: 25,197; Prospective study: 2,500 planned for research-collected data*	~4,300	15,000*	67,639 applicants in round 1 lotteries
Lottery sample size (% of applicants)	2,657 (26%)	4 cohorts with admin data: 5,631 (22%); Prospective study: TBD	~1,900 (44%)	4,500 (30%)	Explore: 172 lotteries (~27.38 children/lottery); Thrive: 36 lotteries (~33.11 children/lottery); Create: 98 lotteries (~28.29 children/lottery)

*Estimated; To be determined

Table 2: Data sources, covariates, and counterfactual data across the five lottery-based preschool studies

Study details	**Boston**	**DC**	**Montessori**	**New Orleans**	**New York City**
Admin. data	X	X	X	X	X
Researcher-collected data	--	X	X	X	X
Covariate data from administrative records	Race/ethnicity, gender, age, home language, free-reduced-priced lunch eligible, country of origin is the U.S.	Gender, age, address, language of application	No administrative data used	Race/ethnicity (for enrollees only), gender, IEP status, SNAP and Medicaid participation, household income, number living in household, home address	Race/ethnicity, age, gender, census tract, Pre-K screening data (for Pre-K enrollees only)
Covariate data from other sources	None (not available)	Additional covariates to be collected via family and educator surveys in the prospective study for ~2,500 students (TBD)	Family survey data: Child birth date, child gender, prior ECE participation, family income, child race/ethnicity	For directly assessed subsample: pretest scores, age in months	For children who apply and enroll in Pre-K (but NOT for children who apply, but don't ultimately enroll): planned to collect baseline child direct assessment data on language, literacy, math, EF, emotion identification, behavior regulation
Counterfactual condition data sources	District and state administrative data (including parent reports from district data)	Family and educator surveys in the prospective study (TBD)	Teacher surveys and observations of a sample of control classrooms	District administrative data contains public program enrollment and parent self-report at kindergarten entry	Same as above. The team only planned to obtain data on children who enrolled in the Pre-K for All system (or NYC school system, for later follow-up)

Table 3: Outcome data across the five lottery-based preschool studies

Study details	Boston	DC	Montessori	New Orleans	New York City
Primary outcomes from administrative data	3rd grade state reading and math standardized test scores	K-3rd grade persistence in public schools, in-grade retention, special education placement, and school and residential mobility, and 3rd grade math and English language arts scores	--	Elementary school application and enrollment behaviors, kindergarten readiness and K-3 literacy scores, attendance, grade retention	--
Exploratory outcomes from administrative data	K-3 school persistence, attendance, receipt of special education services, and grade retention	--	--	--	Third grade test scores, IEP status kindergarten - third grade, attendance kindergarten - third grade
Primary outcomes from other sources	--	Study-collected direct assessments of children's language, literacy, math, social-emotional, and executive function skills at ages 3 and 4	Study-collected direct assessments of children's language, literacy, math, social-emotional, and executive function skills Puzzle task to measure persistence and mastery orientation; Theory of Mind scale.	Study-collected direct assessments of children's literacy, math, working memory, and inhibitory control in fall of Pre-K, kindergarten, and 1st grade	Study-collected direct assessments of language, literacy, math, EF, emotion identification, behavior regulation in Pre-K (Explore primary: math, EF; Thrive primary: emotion ID, behavior reg; Create primary: behavior reg, language)
Exploratory outcomes other sources	--	Study-collected direct assessments of children's racial attitudes (explicit bias) at ages 3 and 4	--	Parent reports of children's socio-emotional wellbeing, parenting stress, and parent-child relationship quality	Child direct assessments: for each PL series, other school readiness outcomes in the list above that weren't identified as "primary" outcomes / targets of PL (e.g., Explore exploratory outcomes were language, literacy, emotion ID, behavior regulation)

Notes: *TBD, data not collected yet.

Table 4. Characteristics of DC three-year-old preschool applicant population, applicants who participated in a lottery, and lottery compliers among applicants in 2014-2018

	All Applicants	Applicants who participated in a lottery	Lottery compliers
Application characteristics			
Num. Schools Ranked	5.59	6.23	7.16
Spanish Application	0.04	0.05	-0.02
Neighborhood characteristics			
Median Income (block-group)	81,341.03	107,240.59	140,836.23
% HS or Less	0.35	0.24	0.12
% Some College	0.20	0.15	0.11
% Bachelors	0.21	0.26	0.31
% Graduate	0.24	0.35	0.46
Population (block)	374.13	304.66	247.11
% Asian	0.03	0.04	0.06
% Black	0.57	0.37	0.19
% Hispanic	0.11	0.13	0.10
% Multi Racial	0.04	0.05	0.06
% White	0.25	0.40	0.58
D.C. Ward of Residence			
Ward 1	0.10	0.15	0.19
Ward 2	0.04	0.07	0.04
Ward 3	0.03	0.06	0.07
Ward 4	0.15	0.21	0.13
Ward 5	0.15	0.14	0.07
Ward 6	0.16	0.22	0.47
Ward 7	0.17	0.08	-0.02
Ward 8	0.20	0.05	0.03
No Ward	0.02	0.02	0.02
Total observations	25,197	5,997	5,997

Source: Author's calculations using My School DC administrative lottery data, OSSE enrollment data, and data from census-type sources reported in a working memo (2022).
Note: Median income and educational attainment is obtained from the 2015-19 ACS estimates at the census block-group level. Population and racial and ethnic shares are derived from the 2020 decennial census population tables at the census block level.