ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-651

Mixed-Delivery Public Prekindergarten: Differences in Demographics, Quality, and Children's Gains in Community-Based versus Public School Programs across Five Large-Scale Systems

Christina Weiland
University of Michigan

Meghan McCormick
MDRC

Jennifer Duer
National Institute for Early Education Research (NIEER)

Allison Friedman-Kraus
National Institute for Early Education Research (NIEER)

Mirjana Pralica
MDRC

Samantha Xia
MDRC

Milagros Nores
National Institute for Early Education Research (NIEER)

Shira Mattera
MDRC

Nearly all states with public prekindergarten programs use mixed-delivery systems, with classrooms in both public schools and community-based settings. However, experts have long raised concerns about systematic inequities by setting within these public systems. We used data from five large-scale such systems that have taken steps to improve equity by setting (Boston, New York City, Seattle, New Jersey, and West Virginia) to conduct the most comprehensive descriptive study of prekindergarten setting differences to date. Our public school sample included 2,395 children in 383 classrooms in 152 schools, while our community-based sample is comprised of 1,541 children in 201 classrooms in 103 community-based organizations (CBOs). We examined how child and teacher demographic characteristics, structural and process quality features, and child gains differed by setting within each of these systems. We found evidence of sorting of children and teachers by setting within each locality, including of children with higher baseline skills and more educated teachers into public schools. Where there were differences in quality and children's gains, these tended to favor public schools. The localities with fewer policy differences by setting – NJ and Seattle – showed fewer differences in quality and child gains. Our findings suggest that inequities by setting are common, appear consequential, and deserve more research and policy attention.

NIEER
NATIONAL INSTITUTE FOR EARLY EDUCATION RESEARCH

VERSION: September 2022

Suggested citation: Weiland, Christina, Meghan McCormick, Jennifer Duer, Allison Friedman-Kraus, Mirjana Pralica, Samantha Xia, Milagros Nores, and Shira Mattera. (2022). Mixed-Delivery Public Prekindergarten: Differences in Demographics, Quality, and Children's Gains in Community-Based versus Public School Programs across Five Large-Scale Systems. (EdWorkingPaper: 22-651). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/pncz-2233

Mixed-Delivery Public Prekindergarten:

Differences in Demographics, Quality, and Children's Gains in Community-Based versus Public School Programs across Five Large-Scale Systems

Christina Weiland[1], Meghan McCormick[2], Jennifer Duer[3], Allison Friedman-Kraus[3], Mirjana Pralica[2], Samantha Xia[2], Milagros Nores[3], & Shira Mattera[2]

[1]University of Michigan

[2]MDRC

[3]National Institute for Early Education Research (NIEER)

Authors' note: The Boston research reported here was conducted as a part of a study funded by R305N160018 – 17 from the Institute of Education Sciences to MDRC. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Thanks to the BPS Department of Early Childhood coaches and staff, the BPS Department of Research, the MDRC team (JoAnn Hsueh, Michelle Maier, Rama Hagos, Jennifer Yeaton, Kelly Terlizzi, Desiree Alderson, Marissa Strassberger, and Sharon Huang), the HGSE research team, and the University of Michigan research team (Lillie Moffett, Paola Guerrero-Rosada, Amanda Weissman, and Christina Daniel). The research from NYC reported here was made possible by a partnership between Robin Hood, one of the country's leading antipoverty organizations based in New York City, and MDRC. Additional funding was provided by the Heising-Simons Foundation, the Overdeck Family Foundation, and the Richard W. Goldman Family Foundation. The New Jersey data used in this work are part of a larger research study funded by the Robert Wood Johnson Foundation (Grant #74677). The authors of this article are solely responsible for the content thereof and shall not be deemed to constitute the position of the funder. Thank you to the districts, schools, centers, teachers, families, and children who worked with us to make this research possible. Thank you also to Charles Whitman, Carol Contreras, and Andrea Kent. The Seattle research was made possible, in part, by the support of the City of Seattle Department of Education and Early Learning. Opinions contained in this article reflect those of the author and do not necessarily reflect those of the City of Seattle Department of Education and Early Learning. Thanks to the City and the schools, centers, teachers, families, and children whose participation made this research possible. The West Virginia research was supported by funding from the West Virginia Department of Education and carried in collaboration with Marshall University. Thanks to Mindy Allenger and to the counties, schools, centers, teachers, families, and children who made this work possible. Thank you to W. Steven Barnett, Ajay Chaudry, and Jason Sachs who provided helpful feedback and comments on a draft of this paper. Correspondence should be sent to Christina Weiland, weilandc@umich.edu.

Abstract

Nearly all states with public prekindergarten programs use mixed-delivery systems, with classrooms in both public schools and community-based settings. However, experts have long raised concerns about systematic inequities by setting within these public systems. We used data from five large-scale such systems that have taken steps to improve equity by setting (Boston, New York City, Seattle, New Jersey, and West Virginia) to conduct the most comprehensive descriptive study of prekindergarten setting differences to date. Our public school sample included 2,395 children in 383 classrooms in 152 schools, while our community-based sample is comprised of 1,541 children in 201 classrooms in 103 community-based organizations (CBOs). We examined how child and teacher demographic characteristics, structural and process quality features, and child gains differed by setting within each of these systems. We found evidence of sorting of children and teachers by setting within each locality, including of children with higher baseline skills and more educated teachers into public schools. Where there were differences in quality and children's gains, these tended to favor public schools. The localities with fewer policy differences by setting – NJ and Seattle – showed fewer differences in quality and child gains. Our findings suggest that inequities by setting are common, appear consequential, and deserve more research and policy attention.

Mixed-Delivery Public Prekindergarten:

Differences in Demographics, Quality, and Children's Gains in Community-Based versus Public School Programs across Five Large-Scale Systems

All but one state with a public prekindergarten program uses a mixed-delivery approach, with seats in public schools and community-based organizations (CBOs). Overall, in 2020-2021, roughly 63% of children in state prekindergarten attended a public school setting and 37% attended a community-based organization (CBO) setting (Friedman-Krauss et al., 2021). Recent policy efforts to expand public prekindergarten suggest that as programs scale-up, the mixed-delivery approach is likely to continue to gain traction. For example, mixed-delivery approaches have characterized preschool expansion efforts in eight large U.S. cities in recent years (Cleveland, Dayton, Denver, New York, San Antonio, San Francisco, Seattle, and Washington D.C.; Karoly, Auger, Kase, McDaniel, & Rademacher, 2016). The Biden Administration's Build Back Better universal preschool proposal for all three and four year olds also emphasized a mixed-delivery approach (White House, 2021).

Despite the popularity of this approach, however, experts have long raised concerns about the structural inequities between public school and CBO settings within many of these systems – i.e., the so-called "two-tier" system problem (Whitebook, 2003). For example, administrators and teachers in many public-school-based prekindergarten settings tend to have higher educational qualifications, often receive higher compensation, and are more likely to be unionized than their CBO counterparts (Bellam, Burton, Whitebook, Broatch, & Young, 2002; Reid, Melvin, Kagan, & Brooks-Gunn, 2019). But with additional scale-up of public prekindergarten on the horizon, more empirical research is needed to better understand the

realities and tradeoffs of mixed-delivery preschool systems. This is particularly important in the context of the COVID-19 pandemic that has had disproportionately negative effects on CBO programs (Weiland et al., 2021).

To help meet this need, we used data from five large-scale mixed-delivery public prekindergarten systems, spanning three large cities (Boston, New York City, Seattle) and two states (New Jersey and West Virginia) to conduct the most comprehensive descriptive study of prekindergarten setting differences to date. We examined how child and teacher demographic characteristics, structural and process quality features, and child gains differed by setting within each system. Importantly, as we detail, included localities had taken explicit steps to address inequities by setting within their mixed-delivery systems. By illuminating inequities that nonetheless persisted pre-COVID, we offer evidence to inform future investments and policies that may be necessary to end the longstanding two-tier problem (Whitebook, 2003). Our findings show that inequities by setting are common, appear consequential, and deserve more research and policy attention.

Mixed-Delivery Systems: Advantages, Disadvantages, and Previous Research

Mixed-delivery approaches have several potential practical advantages, versus implementing programs solely in public schools or CBO settings. First, early education in the U.S. is voluntary and tends to prioritize family choice (Chaudry et al., 2021). A choice in settings can provide families with prekindergarten options that may better meet their needs, preferences, and values. Choice too could increase the likelihood of a cultural, racial/ethnic and/or language match between the family and program staff which can promote children's learning outcomes (Gershenson et al., 2018). Second, having more options for where to

implement prekindergarten programs can ease capacity constraint issues and thus result in additional children served than would otherwise be the case. Third, including a range of providers can mean a larger coalition of supporters in favor of a given prekindergarten initiative, rather than opponents from within the early education community (Ackerman et al., 2009). Finally, tuition for older children in CBOs is often used to offset the higher costs of care for younger children within the same program, as teacher-child ratios are larger for older children. Limiting expansion of public prekindergarten to public schools may hurt the supply and quality of care options for 0-3 year olds (Brown, 2018; Chaudry et al., 2021). These potential benefits help explain why, as we show in Figure 1 in Appendix C, mixed-delivery is the predominant approach used nationally in nearly every state and in many cities to provide state-funded public prekindergarten.

There are also potential downsides of a mixed-delivery approach however. Community-based programs sometimes view public prekindergarten programs, particularly school-based ones, as unwelcome competition which can lead to disagreements over required programmatic elements and standards (Ackerman et al., 2009). CBOs are often underfunded relative to public-school-based programs and/or not held to the same standards (Whitebook, 2003). For example, historically, teachers in public-school-based settings in many systems have significantly higher educational qualifications, higher compensation levels, and lower turnover than in CBOs (Bellam et al., 2002; Johnson et al., 2020). Because CBO teachers are less likely to belong to unions, they generally have less access to critical teaching and learning supports like protected planning time (Yudron et al., 2016).

Despite decades of research on the impacts of preschool on children's early skills, on program structural features, and on classroom processes (Phillips et al., 2017), few empirical studies have paid explicit attention to program setting. Studies examining differences by setting have found empirical support for concerns about the inequities within mixed-delivery systems. For example, a study in New York found substantial differences by setting across structural and programmatic factors (Reid et al., 2019). CBOs offered on average a longer day than public schools (via wrap-around options) and more family services such as mental health services and parenting classes. However, before the introduction of pay parity in New York City, CBO teachers were paid on average about $30,000 annually less than their public school counterparts, while for administrators, the gap was about $53,000. Although public school staff were more highly educated and experienced, almost all CBO teachers and administrators held at least a bachelor's degree. Similar differences have been reported in other systems (e.g., Johnson et al., 2019; Yudron et al., 2016).

These structural differences do seem to result in lower quality experiences for children in CBOs versus their public school counterparts. For example, a study in Boston found that CBO classrooms scored lower on overall quality of interactions and on quality of language and math instruction than public school classrooms (Yudron et al., 2016). A study of math instruction in New York City's mixed-delivery system found that public school prekindergarten programs spent more time on math and had higher quality math instruction than CBO classrooms (McCormick et al., 2021). However, after implementation of a high-quality math curriculum with training and coaching, classrooms in both settings markedly increased time spent on math instruction.

The evidence on child skill gain differences by program type/setting is even more limited and so far, is nuanced. In Georgia's mixed-delivery system, children in public-school-based programs showed larger gains in language, math, and socio-emotional skills than their counterparts in CBOs (Peisner-Feinberg et al., 2019). However, in the aforementioned New York City study of the implementation of a high-quality math curriculum, there were impacts on children's math skills only in the public schools and on children's vocabulary and executive function only in CBOs (McCormick et al., 2021).

In considering the evidence from these mostly descriptive studies, it is important to note too that the children differed markedly by setting within mixed-delivery systems – i.e., there was selection into setting types. Compared to children in public school programs, students enrolled in CBOs are more likely to be from families with low incomes and from a racial/ethnic or language minority group (Burchinal et al., 2008; Reid et al., 2019). At least some of the time, these differences are by design. The Boston program in the first years of its mixed-delivery system, for instance, intentionally expanded to CBOs in neighborhoods with higher rates of poverty and lower rates of preschool program enrollment (Yudron et al., 2016). But the limited research available suggests selection does not appear to entirely explain public school and CBO differences. For example, research using data from the Head Start Family and Child Experiences Survey (FACES) found that Head Start programs located in public school settings have higher levels of instructional support and classroom organization than programs in CBOs, even though programs in both settings primarily enrolled students experiencing poverty (Alamillo et al., 2015).

Current Study and Research Questions

Our study builds directly from the previous empirical literature on differences by setting in mixed-delivery systems. Using data from mixed-delivery preschool programs in three large cities (Boston, New York City, and Seattle) and two states (New Jersey and West Virginia), we address three research questions:

1) How do public prekindergarten programs differ in their child and teacher characteristics in public school versus CBO settings?

2) How do structural and process quality differ in public school versus CBO prekindergarten settings?

3) How do children's gains in literacy, mathematics, socio-emotional, and executive function skills differ in public school versus CBO prekindergarten settings?

Method

Sample details

Within each of the five localities – Boston, New York City, Seattle, New Jersey, and West Virginia – children in public schools and CBOs were recruited at the same time. See Table 1 for Ns for children, classroom, schools, and CBO by locality. In all, our public school sample includes 2,237 children in 366 classrooms in 146 schools. Our CBO sample is comprised of 1,707 children in 221 classrooms in 110 CBOs. For parsimony and because sampling details have been detailed previously (see Nores & Contreras, 2021; Nores, Friedman-Kraus, & Figueras-Daniel, 2022; Nores et al., 2019; Morris, Mattera, & Maier, 2016; McCormick et al., 2020), full details on sample selection processes by locality are in Appendix A. The sampling strategy varied by locality from inclusion of all sites in the program (Seattle) to a selection of sites that met particular criteria (NYC). Accordingly, the representativeness of findings from each locality varies – a point we return to our limitations sections.

Procedures

Data collection in each locality was approved by the relevant Institutional Review Board. For parsimony, we refer the reader to full details on procedures in Appendix A. In brief, all five study teams actively consented participating sites, teachers, families, and children; children were assessed in pull-out sessions by a trained child assessor; and classrooms were videotaped (Boston) or observed in-person (all others), for scoring by a trained and reliable coder on the CLASS measure (Pianta et al., 2008). Observational data in all sites demonstrated high levels of interrater reliability. Data collectors assessed children in the fall and spring of prekindergarten, with the exception of WV, where children were assessed in fall of prekindergarten and fall of kindergarten. We control for the time between initial and follow-up testing in child gains models and also return to this timing difference in our limitations section.

Measures

We outline our key study measures below. For parsimony, additional measures details are in Appendix B.

Classroom process quality. In each locality, we measured classroom process quality using the Classroom Assessment Scoring System (CLASS) PreK (Pianta, La Paro, & Hamre, 2008). This observational tool measures three domains of teacher-child interactions: Emotional Support, Classroom Organization, and Instructional Support. All dimensions are scored on a seven-point scale, with higher scores indicating higher quality.

Vocabulary. In all localities but New York City, we used the Peabody Picture Vocabulary Test-IV to measure children's receptive vocabulary in standard American English. Consistent with other preschool studies that have used the PPVT (e.g., Weiland & Yoshikawa, 2013), we used the total raw score as our outcome measure.

In New York City, we assessed children's receptive language using the Receptive One Word Picture Vocabulary Test, 4th edition (ROWPVT-4; Martin & Brownell, 2011). In analyses, we used the raw score to be consistent with our PPVT decision in other localities.

Math. To assess children's early math skills, we used the Woodcock–Johnson Applied Problems III (Mather, McGrew, & Wendling, 2001) subtest in all localities. Consistent with prior studies (Weiland & Yoshikawa, 2013), we used the raw score in our analysis. In NYC only, we used the Early Childhood Longitudinal Study–Birth Cohort (ECLS-B) math assessment in the fall (Najarian et al., 2010).

Executive function. In all sites but Boston, we used the Pencil Tap task (Diamond & Taylor, 1996) to measure inhibitory control. We used the proportion of trials (out of 16) that a child got correct to operationalize the score. In Boston only, we used the Hearts and Flowers task to measure inhibitory control (Davidson, Amso, Anderson, & Diamond, 2006). We used the Incongruent Trial score because it best approximates the same construct as the Pencil Tap measure (Diamond, Barnett, Thomas, & Munro, 2007) that was collected in other localities.

Child and family demographics. From administrative data and/or parent surveys (see Appendix A Procedures), we created variables capturing child demographics for gender, age in years as of September 1 of prekindergarten fall, race/ethnicity (Asian, Black, Latino, White, or Other), Individual Education Plan (IEP) status, and home language (English or other) for all localities. For Boston, NYC, and NJ, we also created a variable for eligibility for free-or-reduced price lunch and in WV, an indicator that the child was from a family with low income, as defined in their system. For Boston, NYC, and NJ, we also created variables capturing parent education (less than high school, high school/GED, some college, or BA+).

Teacher and classroom variables. We used survey and administrative data to capture teacher gender, race/ethnicity (same categories as for children), education (less than BA, BA, MA+), whether the teachers' highest degree was in early childhood education (ECE), whether the teacher held any state teaching certification and whether it was in ECE, total years of teaching experience, total years teaching at current school, years teaching prekindergarten, and language(s) spoken (speaks Spanish and/or a non-Spanish/non-English language). All these characteristics were available in Boston, nearly all were in NYC (one missing), and some in Seattle, NJ, and WV (see Table 3). We also collected information on class size (all localities) and ratios (available in Boston, NYC only) using administrative data, teacher reports, and classroom observations.

Data Analysis

For research questions 1 and 2 – setting differences in child and teacher characteristics; structural quality; and instructional quality – we used simple descriptive statistics to calculate the mean and standard deviation for public schools versus CBOs. We used independent sample two-tailed t-tests to examine whether differences in means by setting type within each locality were statistically significantly different from one another.

For research question 3 – difference in child gains by setting – we used residualized gain models with random intercepts for classrooms. We regressed each skill of interest at the end of prekindergarten (or beginning of K in WV) on an indicator for whether the child attended a public school or CBO. These models also included covariates for the corresponding fall of prekindergarten skill and child and family characteristics described in the previous section that have predicted children's outcomes in other studies (e.g., Bloom & Weiland, 2015), covering

child gender, age, race/ethnicity, home language, IEP status, and time between fall and spring testing for all localities. We also included parent education (Boston and NYC) and family income (Boston, NYC, NJ, and WV). In NJ and WV, we also included fixed intercepts for district; findings from models with alternative specifications for these two states are in our robustness checks section.

Due to data restrictions, we conducted all analyses separately within locality – i.e., we were unable to pool data to produce joint estimates. We examine patterns and magnitudes of setting differences, in addition to statistical significance.

Data were missing at a low rate (e.g., Boston 0-6%), with a few exceptions (e.g., WV was missing 20% for gender, NYC planned missingness of 38% for baseline scores due to resource constraints). For covariates, we used a missing data dummy strategy (e.g., Bloom & Weiland, 2015) to retain cases and adjust estimates for missingness. For missing outcome data, we used listwise deletion following Graham (2009).

Results

Policy differences by setting. Before presenting results, we highlight some key policy elements of these localities' mixed-delivery systems that are critical for interpretation. Overall, all localities took steps to try to ensure equity across settings. There was nonetheless evidence of policy differences by setting in three localities (Boston, NYC, and WV) and less so in NJ and Seattle.

As shown in Table 1, the included localities range from long-established (WV, 1983) to recently developed (Seattle, 2005) programs. All but Boston were mixed-delivery from the start;

Boston's program was public-school-based only for its first six years. In the current study years, Boston and NYC CBOs were subject to different standards than public schools; in other localities, the standards were the same by setting. In terms of salary, three localities had parity between CBOs and public schools (and between prekindergarten and K-12), though in one of these (Boston), CBO teachers may have had to work a longer year (varied by CBO). All required at least a BA in both CBO and public school settings; Boston required a masters degree within five years in public schools but not CBOs. Curriculum requirements did not vary by setting in any locality, though in Boston, all prekindergarten classrooms were expected to use the same curriculum (McCormick et al., 2021). In Seattle, NJ, and WV, sites chose from a list of curricula and in NYC, there was no such list for CBOS or public schools.

RQ 1: Child and teacher characteristics. Public schools and CBOs served different child and family populations (Table 2). Within the five localities, children in public school-based programs had higher baseline test scores than their CBO peers for 12/15 comparisons. Children in CBOs scored higher for two comparisons; for inhibitory control in NYC, there was no difference. Eight baseline skill differences were statistically significant ($p<.05$, all favoring public schools). Using CBO standard deviations, these statistically significant differences ranged from 0.15 SDs (Seattle language) to 0.61 SDs (Boston math). Interpreted in months of learning, these significant differences ranged from a 1 month of language learning in Seattle to 6.5 months of math in Boston (Hill, Bloom, Black, & Lipsey, 2008).

In terms of race/ethnicity and home language, localities overall served very different student populations that reflected differences in location demographics; WV, for example, served mostly white, English home language students while NJ and NYC samples were majority Latino.

In all localities but NYC, public schools enrolled proportionately more White students than CBOs (range of 9-25 percentage points; see Table 2). CBOs in three localities (NYC, Seattle, and NJ) enrolled higher proportions of Latino students than their public school counterparts. Black students were more concentrated in CBOs compared to public schools in all localities but NYC. For home language, CBOs enrolled more DLLs in NYC, NJ, and Seattle, while public schools enrolled proportionately more DLLs in Boston and the split was about equal in WV. Children with more educated parents were concentrated in public schools in two of the three localities with parent education information available (Boston and NYC), with no such differences in NJ. For family income information, children eligible for free-or-reduced priced lunch were concentrated in higher numbers in CBOs compared to public schools in Boston, NYC, and NJ and were about equally split between settings in WV.

Masters-level teachers were much more likely to work in public school than CBO settings in all localities except NJ (Table 3), with differences ranging from 8 percentage points (WV) to 51 percentage points (Seattle). In all four sites with teacher race/ethnicity data available, the most pronounced difference was for White teachers; White teachers were much more likely to work in public school than CBO settings (17-44 percentage point differences across localities). Teachers in Boston, NJ, and WV were about equally as experienced by setting; in NYC, in contrast, CBO teachers had about 11 years of experience teaching versus 18 for public school teachers.

RQ 2: Structural and process quality. As shown in Table 4, differences in class size and teacher-child ratios by setting were small in magnitude, only one was statistically significant and did not consistently favor one setting across localities.

For process quality, across localities, public school classrooms scored higher than CBO classrooms on 12/15 comparisons, though only three of these differences were statistically significant ($p<.05$). One difference favoring CBOs in NJ was statistically significant (Instructional Support, $p<.05$). Magnitudes of the 12 differences that favored public schools ranged from 0.09 SDs (NYC Instructional Support) to 1.02 SDs (Boston Classroom Organization). Magnitudes for the three differences favoring CBOs ranged from 0.23 (NYC Classroom Organization) to 0.53 (Boston Classroom Organization) SDs.

RQ 3: Child gains. On average, children made substantial gains in their skills from fall to spring (Table 1). However, as shown in Figure 1, children in public schools made statistically significantly larger gains in their language skills than their CBO counterparts of about 0.18 SDs in both NYC ($p<0.5$) and WV ($p<.10$). In Boston, children in public school settings made larger gains in their math skills than their CBO counterparts of about 0.26 SDs ($p<0.5$). No other tested differences by setting were statistically significant, although the CBO coefficient was negative in magnitude in 10/15 comparisons, ranging between 0.03 to 0.26 SDs. The negative association between CBO and child gains was larger than 0.10 SDs for 5/15 comparisons. For the five positive associations for enrolling in CBOs, magnitudes ranged between 0.02 to 0.07 SDs (for Boston language and NYC inhibitory control, respectively).

Robustness checks. We examined whether NJ and WV child gains results were sensitive to how we chose to model clustering within districts. As shown in Appendix Table 2, results were similar across specifications for NJ. For WV, magnitudes were very similar but the CBO difference for vocabulary was statistically significant at the 0.05 level in the random intercept approach (vs. $p=0.06$ in the primary specification).

We also examined the sensitivity of our decision to code the small number of Head Start programs located in public schools as community-based programs using NJ data (the locality with a larger relative number of such sites). We instead dropped such sites and refit child gains models. We also tried coding them as public and then refit child gains models. In these alternatives, magnitudes of differences were larger, changed direction for inhibitory control, and became statistically significant for math in one case (Appendix Table 3). But overall, the interpretation of our findings for NJ as showing fewer difference by setting relative to other localities was unchanged.

Finally, we examined the possibility that children in CBOs may have made larger gains in other important domains than their public school counterparts due to differential emphasis in skill type by setting. We examined gains in social-emotional skills in the three localities with available data (Boston, NYC, and NJ). Findings are provided in Appendix C. For these localities, there was no evidence that children in CBOs were making more gains than their public school counterparts in this domain.

Discussion

We examined differences in CBO versus public school prekindergarten settings within five large-scale mixed-delivery public prekindergarten systems – Boston, New York City, Seattle, New Jersey, and West Virginia – to conduct the most comprehensive descriptive study of prekindergarten setting differences to date. Our findings suggest some setting differences in child and teacher characteristics; differences in process but not structural quality; and in child gains. When differences were found, they generally favored public school settings over CBO settings, even though each of the localities had taken steps to explicitly address inequities within

their mixed-delivery systems. Our findings, which we discuss in turn below, offer potential insights into the additional investments and policies that may be necessary to improve equity in mixed-delivery systems.

First, we found some evidence of sorting into CBOs versus public schools at the student level in terms of child skills and race/ethnicity and home language. Across localities, CBOs generally served children with lower baseline language and math than their public school counterparts. This selection into setting meant that on average, children in CBOs experienced peers with lower average skills than children in public schools. Notably, some correlational work shows that having peers with stronger language skills may support children's skill development in preschool (Henry & Rickman, 2007). Lower peer skills in CBOs than in public schools may be a contributing factor to smaller gains in children's language and math skills we found in CBOs relative to public school prekindergarten in a subset of localities.

The most consistent difference in child demographics by setting across localities was that in all but one locality (NYC), public schools served proportionately more White children than CBOs. Similarly, for teachers, we also found that White teachers were much more likely to work in public school than CBO settings. These findings have nuanced implications. Previous research has shown that ECE settings are more racially segregated overall in the U.S. than K-12 (Greenberg et al., 2019) and that racial integration has benefits for older children (Johnson, 2019). Research has also pointed to the inequities and racialization of children's early learning experiences in segregated classrooms (Adair & Colegrove, 2021). But on the other hand, teacher-student race match has been shown to promote the learning of Black and Latino students in particular as early as kindergarten (e.g., Gershenson et al., 2018), though research in preschool is

more scant. Accordingly, the representation of Black and Latino teachers in CBO prekindergarten settings, coupled with the fact that CBOs tend to enroll children of color in higher proportion than public schools, suggests a strength of CBOs to preserve and replicate within public school prekindergarten programs. But at the same time, racial segregation may have negative implications for students' development and experiences. More research is needed on the prevalence, systemic reasons for, and consequences of racial/ethnic segregation in early childhood settings and systems, as well as specific mechanisms for promoting racial integration.

Our finding that teachers in public schools were more educated than their CBO counterparts (except in NJ) suggests an interplay between policy and selection. In Boston, for example which had 50 percentage point difference in masters-level teachers by setting, there is a policy difference at play – only teachers in public schools and not CBOs were required to have a masters degree within five years. But all other localities had the same requirements by setting and still, all localities but NJ also showed disparities in masters degree teachers favoring public schools. More equity in policies by setting, along with MA degree pathways programs specifically targeted to teachers in CBOs and careful consideration of when exceptions may be allowed (Friedman-Kraus et al., 2021), are possible approaches to addressing these disparities.

Our quality and child gains findings underscore the importance of more equity-focused work in mixed-delivery systems. Importantly, in all localities and settings, children made gains from fall to spring. But public school classrooms outscored CBO classrooms on 12/15 comparisons across the CLASS subscales. Children in CBOs generally made lower gains on their language, math, and inhibitory control skills than their public school counterparts. We found this pattern despite the fact that children in CBOs began the year with considerably lower

skills. Typically, children with "more room to grow" experience larger gains in public preschool programs (Bloom & Weiland, 2015; Phillips et al., 2017). Accordingly, these findings are troubling, especially since prekindergarten programs are often identified as a potential lever for addressing racial and income-based opportunity gaps (Chaudry et al., 2021). Notably, differences in child gains by setting were smaller in New Jersey and Seattle, the two systems that our policy summary (Table 1) showed to be the most equitable by setting among the five localities. Together, these findings point to the importance of policies and supports to enhance equity across settings.

Our work has several key limitations. This study is descriptive and correlational in nature. Teachers and children were not randomly assigned to setting types. Due to data sharing restrictions, we also could not pool the data; some within-locality analyses may be under-powered. Accordingly, we tried to interpret patterns and magnitudes across sites and not rely solely on statistical significance to understand results. Better-powered studies across multiple localities are needed. We also sampled in each locality from the broader population of CBO and public school teachers in all localities but Seattle; data on the full populations are not available. The sample in NYC was specifically recruited from programs primarily serving students from families with lower-incomes that had also agreed to participate in a larger study of a prekindergarten math curriculum (see Appendix A). We therefore cannot describe the full extent to which our results may generalize to each locality and setting. We selected localities based on a combination of data to which we had access and with the goal of including localities that had taken steps to address equity by setting. Our findings accordingly do not apply to all mixed-delivery prekindergarten systems in the U.S. Relatedly, our study did not include family child care homes which are sometimes also part of mixed-delivery preschool systems. In terms of

measurement, we did not have common measures of other important child developmental domains for all sites such as literacy and socio-emotional skills. Our measures of structural and process quality too were limited. Though used widely and often integrated into early learning accountability systems, our process quality measure – the CLASS – has been shown to have null or small predictive validity with children's gains (Guerrero Rosada et al., 2022). Outcome measures in West Virginia are from fall of kindergarten versus spring of prekindergarten for other localities, although we did control for time between assessments in our child gains models. In NYC, our math outcome and baseline measures were from different assessments. We also lacked measures of student attendance, another potential driver of differences in children's gains that might differ by setting (Arbour et al., 2017). Finally, we lacked data on more fine-grained program characteristics like protected planning time for teachers, curriculum fidelity, coaching specifics, director and principal characteristics, and special education supports. Such data should be prioritized in future investigations of differences by setting in mixed-delivery systems.

Taken together, our descriptive findings suggest that inequities by setting are common, appear consequential for children's learning, and deserve more research and policy attention. Providing equitable, high-quality early learning in mixed-delivery systems requires intentional policies, additional investments, and careful monitoring and research, with particular attention to programs serving historically marginalized communities.

References

Ackerman, D. J., Barnett, W. S., Hawkinson, L.E., Brown, K. & McGonigle, E.A. (2009). *Providing preschool education for all 4-year-olds: Lessons from six state journeys*. NIEER Preschool Policy Brief. http://nieer.org/wp-content/uploads/2016/08/19-1.pdf

Adair, J. K., & Colegrove, K. S. S. (2021). *Segregation by experience: Agency, racism, and learning in the early grades*. University of Chicago Press.

Alamillo, J., Aikens, N., Moiduddin, E., Bush, C., Malone, L., & Tarullo, L. (2018). *Head Start programs in spring 2015: Structure, staff, and supports for quality from FACES 2014*. OPRE Report, 79. https://www.acf.hhs.gov/opre/report/head-start-programs-spring-2015-structure-staff-and-supports-quality-faces-2014

Arbour, M., Yoshikawa, H., Willett, J., Weiland, C., Snow, C., Mendive, S., ... & Treviño, E. (2016). Experimental impacts of a preschool intervention in Chile on children's language outcomes: Moderation by student absenteeism. *Journal of Research on Educational Effectiveness*, *9*, 117-149.

Barrueco, S., Lopez, M., Ong, C., & Lozano, P. (2012). *Assessing Spanish-English bilingual preschoolers: A guide to best approaches and measures*. Paul H Brookes Publishing.

Bellam, D., Burton, A., Whitebook, M., Broatch, L., & Young, M. P. (2002). *Inside the pre-K classroom: A study of staffing and stability in state-funded prekindergarten programs*. Center for the Child Care Workforce. https://cscce.berkeley.edu/inside-the-pre-k-classroom-a-study-of-staffing-and-stability-in-state-funded-pre-k-programs/

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, NY: MDRC. https://files.eric.ed.gov/fulltext/ED558509.pdf

Brown, J. H. (2018). Does public pre-k have unintended consequences on the child care market for infants and toddlers? *Princeton University Industrial Relations Section Working Paper*, *626*.

Burchinal, M., Nelson, L., Carlson, M., & Brooks-Gunn, J. (2008). Neighborhood characteristics, and child care type and quality. *Early Education and Development*, *19*(5), 702-725.

Chaudry, A., Morrisey, T., Weiland, C., & Yoshikawa, H. (2021). *Cradle to kindergarten: A new plan to combat inequality* (2nd ed). New York, NY: Russell Sage Foundation.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037-2078.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". *Developmental Psychobiology, 29*(4), 315–334.

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science, 318* (5855), 1387–1388. http://dx.doi.org/ 10.1126/science.1151148

Duncan, S., & DeAvila, E. (1998). *PreLAS 2000 Technical Report*. Monterey, CA: CTB⁄ McGraw Hill.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test*. Pearson Assessments.

Frede, E., & Barnett, W. S. (2011). Why pre-k is critical to closing the achievement gap. *Principal*, *90*(5), 8-11.

Frede, E., Jung, K., Barnett, W. S., Lamy, C. E., & Figueras, A. (2007). *The Abbott preschool program longitudinal effects study (APPLES)*. New Brunswick, NJ: National Institute for Early Education Research.

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G. G. & Gardiner, B. A. (2021). *The State of Preschool 2020: State Preschool Yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

Gershenson, S., Hart, C. M., Hyman, J., Lindsay, C., & Papageorge, N. W. (2018). *The long-run impacts of same-race teachers* (No. w25254). National Bureau of Economic Research. https://doi.org/10.3386/w25254

Goldman, S. R., Greenleaf, C., Yukhymenko-Lescroart, M., Brown, W., Ko, M. L. M., Emig, J. M., ... & Britt, M. A. (2019). Explanatory modeling in science through text-based investigation: Testing the efficacy of the Project READI intervention approach. *American Educational Research Journal*, *56*(4), 1148-1216.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Greenberg, E., Monarrez, T., Feng, A., Feldman, A., Hinson, D., & Peiffer, E. (2019). *Segregated from the start: Comparing segregation in early childhood and K-12 education.* https://www.urban.org/features/segregated-start

Gresham, F. M., Elliott, S. N., Vance, M. J., & Cook, C. R. (2011). Comparability of the Social Skills Rating System to the Social Skills Improvement System: Content and psychometric comparisons across elementary and secondary age levels. *School Psychology Quarterly, 26*(1), 27–44. https://doi.org/10.1037/a0022662

Guerrero-Rosada, P., Weiland, C., McCormick, M., Hsueh, J., Sachs, J., Snow, C., & Maier, M. (2021). Null relations between CLASS scores and gains in children's language, math, and executive function skills: A replication and extension study. *Early Childhood Research Quarterly*, *54*, 1-12.

Henry, G. T., & Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review, 26*(1), 100–112.

Hightower, A. D., Cowen, E. L., Spinell, A. P., Lotyczewski, B. S., Guare, J. C., Rohrbeck, C. A., & Brown, L. P. (1987). The Child Rating Scale: The development of a socioemotional self-rating scale for elementary school children. *School Psychology Review, 16*(2), 239–255. https://doi.org/10.1080/02796015.1987.12085288

Hilbert, S., Nakagawa, T. T., Bindl, M., & Bühner, M. (2014). The spatial Stroop effect: A comparison of color-word and position-word interference. *Psychonomic Bulletin & Review, 21*(6), 1509–1515. http://dx.doi.org/10.3758/ s13423-014-0631-4

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

Johnson, R. C. (2019). *Children of the dream: Why school integration works*. Basic Books.

Johnson, A. D., Martin, A., & Schochet, O. N. (2019). How do early care and education workforce and classroom characteristics differ between subsidized centers and available center-based alternatives for low-income children? *Children and Youth Services Review*, *107*. https://doi.org/10.1016/j.childyouth.2019.104567

Karoly, L. A., Auger, A., Kase, C. A., McDaniel, R. C., & Rademacher, E. W. (2016). *Options for investing in access to high-quality preschool in Cincinnati*. Santa Monica, CA: RAND. https://doi.org/10.7249/RR1615

Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., & Wilson, S. J. (2017). Learning-related cognitive self-regulation measures for prekindergarten children: A comparative evaluation of the educational relevance of selected measures. *Journal of Educational Psychology*, *109*(8), 1084.

Martin, N.A., & Brownell, R. (2011). *Expressive one-word picture vocabulary test-4 (EOWPVT-4).* Novato, CA: Academic Therapy Publications.

McCormick, M. P., Mattera, S. K., Maier, M. F., Xia, S., Jacob, R., & Morris, P. A. (2022). Different settings, different patterns of impacts: Effects of a Pre-K math intervention in a mixed-delivery system. *Early Childhood Research Quarterly*, *58*, 136-154.

McCormick, M., Weiland, C., Hsueh, J., Pralica, M., Weissman, A. K., Moffett, L., ... & Sachs, J. (2021). Is skill type the key to the prek fadeout puzzle? Differential associations between enrollment in preK and constrained and unconstrained skills across kindergarten. *Child Development*, *92*(4), e599-e620. https://doi.org/10.1111/cdev.13520

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

Najarian, M., Snow, K., Lennon, J., Kinsey, S., & Mulligan, G. (2010). *Early childhood longitudinal study, birth cohort (ECLS-B), preschool–kindergarten 2007 psychometric report (NCES 2010-009).* National Center for Education Statistics, Institute of Education Sciences, US Department of Education, Washington, DC.

Nores, M., Barnett, S., Jung, K., Joseph, G., & Bachman, L. (2019). *Year 4 Report: The Seattle Preschool Program*. New Brunswick, NJ: National Institute for Early Education Research & Seattle, WA: Cultivate Learning.

Nores, M., & Contreras, C. (2021). *Evaluation of West Virginia Universal Pre-K: Summary report*. New Brunswick, NJ: National Institute for Early Education Research.

Nores, M., Friedman-Krauss, A., & Figueras-Daniel, A. (2022). Activity settings, content, and pedagogical strategies in preschool classrooms: Do these influence the interactions we observe? *Early Childhood Research Quarterly*, *58*, 264-277.

Peisner-Feinberg, E., Van Manen, K., Mokrova, I., & Burchinal, M. (2019). *Children's outcomes through second grade: Findings from Year 4 of Georgia's Pre-K Longitudinal Study*. FPG Child Development Institute.

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. https://www.brookings.edu/wp-content/uploads/2017/04/consensus-statement_final.pdf

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system: Manual K-3*. Baltimore, MD, US: Paul H Brookes Publishing.

Puma, M., Bell, S., Cook, R., Heid, C., & U.S Department of Health and Human Services. (2010). *Head Start Impact Study Final Report.* Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Reid, J. L., Melvin, S. A., Kagan, S. L., & Brooks-Gunn, J. (2019). Building a unified system for universal Pre-K: The case of New York City. *Children and Youth Services Review*, *100*, 191-205. https://doi.org/10.1016/j.childyouth.2019.02.030

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22*(2), 173–187. https://doi.org/10.1016/j.ecresq.2007.01.002

Weiland, C. (2018). Commentary: Pivoting to the “how”: Moving preschool policy, practice, and research forward. *Early Childhood Research Quarterly*, *45*, 188-192. https://doi.org/10.1016/j.ecresq.2018.02.017

Weiland, C., Greenberg, E., Bassok, D., Markowitz, A., Guerrero Rosada, P. … & 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.* Ann Arbor, MI and DC: University of Michigan Education Policy Initiative and Urban Institute Policy Brief. https://edpolicy.umich.edu/files/EPI-UI-Covid%20Synthesis%20Brief%20June%202021.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.

Whitebook, M. (2003). *Early education quality: Higher teacher qualifications for better learning environments—A review of the literature*. Report for the Center for the Study of Child Care Employment. Berkeley, CA.

White House. (2021, October 28). *Build Back Better Framework*. Retrieved from https://www.whitehouse.gov/briefing-room/statements-releases/2021/10/28/build-back-better-framework/

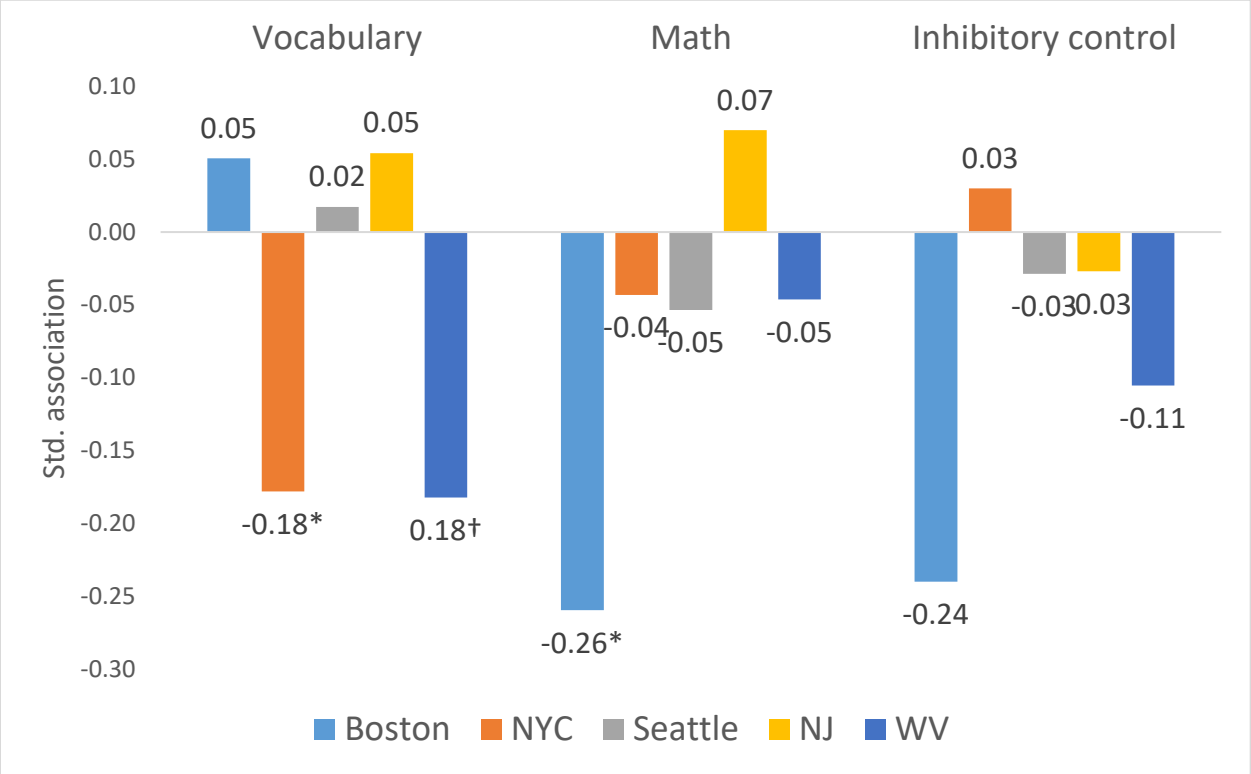
Woodcock, R. W., Mather, N., McGrew, K. S., & Wendling, B. J. (2001). *Woodcock–Johnson III tests of cognitive abilities*. Itasca, IL: Riverside Publishing Company.

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology, 5*. http://dx. doi.org/10.3389/fpsyg.2014.00213

Yudron, M., Weiland, C., & Sachs, J. (2016). *BPS K1DS: Piloting the Boston Public Schools' prekindergarten model in community-based organizations*. Boston, MA: Boston Public Schools.

Tables and Figures

Figure 1: Standardized associations between attending a CBO site (versus public school) and children's gains in public prekindergarten



Note: *p<0.05; †p<0.10. Models included covariates for the corresponding fall of prekindergarten skill and a host of child and family characteristics (child gender, age, race/ethnicity, home language, IEP status, and time between fall and spring testing for all localities; parent education in Boston and NYC; and family income in Boston, NYC, NJ, and WV). In all sites, we included random intercepts for classroom and in NJ and WV, we also included fixed intercepts for district.

Table 1: Sample and program details

	Boston	NYC	New Jersey	West Virginia	Seattle
Sample Ns & timing					
Public schools	20	23	47	45	11
Classrooms	42	61	180	64	19
Children	307	491	732	401	306
CBOs	10	11	35	16	38
Classrooms	10	25	97	33	56
Children	79	200	387	205	836
Data collection timing	2016-2017	2014-2015	2017-2018 & 2018-2019	Fall 2015 & Fall 2016	2018-2019
Program details					
Year program started	2005-2006	2014-2015	1998	1983	2015-2016
Mixed-delivery history	Public schools-only until 2012	Mixed delivery from the start	Mixed delivery from the start	Commitment to 50% mixed delivery began in 2002	Mixed delivery from the start
Same standards by setting? *	State: Yes; Local: No	State: Yes; Local: No~	Yes	Yes	Yes
Public school K-12 pay parity	Yes	Yes	Yes	Yes	Yes
CBO K-12 pay parity **	Partial	No	Yes	No	Yes
Teacher qualification requirements	Public schools: BA min., Master's within 5 years; CBOs: BA min.	BA	BA in ECE	BA in ECE/CD/ECE Sped	BA in ECE
Funding model	Public schools: City funded; CBOs braided	Braided/blended with Head Start, state, and local resources	State aid formula, with different amounts for	State-aid funding for pre-K is allocated as part of the	City property tax levy

	across federal, state, local, and private pay		public schools, child care, Head Start	K-12 state aid funding process	
Hours	6.5 hours/day	6 hours/day	6 hours/day	Min. 14 hours per week***	6 hours/day
N days per week	5	5	5	varies	5
Curriculum requirements	Boston Focus Curriculum	No requirements	The Creative Curriculum, HighScope Preschool Curriculum, Tools of the Mind, Connect4Learning	Required to use one of: Creative Curriculum; High Reach; HighScope	HighScope or Creative Curriculum
Student assessment requirements	Public Schools: PALS; CBOs - Teaching Strategies Gold (if accepting state funding)	No requirements	Programs may use: Early Learning Scale; HighScope COR; Teaching Strategies GOLD; Work Sampling	Early Learning Scale (ELS) (multiple times during the year)	Teaching Strategies Gold
Other quality monitoring	Periodic observational measures (of overall quality and curriculum fidelity)	Varied by program model; CBOs offered more supports than public schools	ECERS, Self-Assessment and Validation System (SAVS)	WV UPK Health and Safety Checklist; annual observations overseen by each county (tools determined locally).	CLASS

* Boston CBO sites followed state requirements for ratios, group size, and health/safety, while public school sites followed BPS standards. For NYC, all sites followed the same state standards but local standards differed by site type.
** Boston: Parity with K-12 in public school-based prekindergarten; in community-based sites, teachers with a BA had a minimum of the entry-level salary of a public school teacher. NYC: Parity with K in public schools; no parity at this time in pay between CBOs and public schools (changed in later years);

***The minimum requirement has changed since the time WV data were collected and is now 1,500 minutes per week (25 hours).

Table 2: Child baseline measures, outcome measures, and demographics by locality and setting

	Boston				New York City				Seattle				New Jersey				West Virginia			
	Public	CBO	Diff.		Public	CBO	Diff.		Public	CBO	Diff.		Public	CBO	Diff.		Public	CBO	Diff.	
Panel 1: Child skills																				
Language																				
Baseline	73.27	67.43	5.84	†	44.89	43.86	1.03		71.16	64.05	7.11	***	54.03	47.22	6.68	***	84.27	74.85	9.41	**
Follow-up	87.79	78.58	9.21	**	54.99	50.05	4.94	**	83.41	75.93	7.48	***	66.83	61.93	5.29	**	110.84	105.38	5.46	**
Math																				
Baseline	12.52	9.75	2.77	***	19.06	20.46	-1.40	†	10.74	9.93	0.81	*	7.16	6.33	0.73	**	17.00	15.67	1.33	***
Follow-up	15.86	11.88	3.98	***	12.06	11.86	0.21		14.05	13.02	1.03	**	10.06	9.38	0.59	*	20.81	19.79	1.02	**
Inhibitory Control																				
Baseline	0.59	0.55	0.04	*	0.50	0.51	0.00		0.42	0.38	0.04		0.28	0.29	-0.01		0.82	0.81	0.01	
Follow-up	0.68	0.58	0.10	***	0.69	0.70	-0.01		0.57	0.52	0.05	†	0.48	0.43	0.05	*	0.90	0.88	0.02	
Panel 2: Demographics																				
Female	0.51	0.51	0.00		0.51	0.56	-0.05		0.50	0.48	0.02		0.49	0.52	-0.03		0.49	0.51	-0.02	
Age (in years)	4.51	4.46	0.05		4.18	4.18	0.00		4.32	4.15	0.17	***	4.12	3.96	0.18	***	4.47	4.43	0.04	
Race/ethnicity																				
Asian	0.16	0.06	0.10	**	0.02	0.03	-0.01		0.20	0.17	0.03		0.03	0.04	-0.01		0.01	0.00	0.01	
Black	0.21	0.69	-0.48	***	0.37	0.31	0.06		0.20	0.27	-0.07	*	0.15	0.19	-0.04		0.03	0.09	-0.06	***
Latino	0.30	0.22	0.08		0.58	0.63	-0.05		0.10	0.16	-0.06	*	0.51	0.64	-0.13	***	0.01	0.00	0.01	
White	0.28	0.03	0.25	***	0.01	0.02	-0.01		0.30	0.19	0.11	***	0.28	0.10	0.18	***	0.93	0.84	0.09	***
Other	0.06	0.00	0.06	***	0.02	0.01	0.01		0.16	0.17	-0.01		0.03	0.03	0.00		0.04	0.05	-0.01	
Parent ed. Level																				
<HS	0.11	0.16	-0.05		0.27	0.33	-0.06		--	--	--	--	0.13	0.17	-0.04		--	--	--	--
HS/GED	0.19	0.16	0.03		0.30	0.27	0.03		--	--	--	--	0.35	0.33	0.02		--	--	--	--
Some College	0.25	0.49	-0.24	***	0.28	0.28	0.00		--	--	--	--	0.32	0.29	0.03		--	--	--	--
BA+	0.45	0.18	0.27	***	0.14	0.12	0.03		--	--	--	--	0.19	0.2	-0.01		--	--	--	--
Non-Eng HL	0.54	0.27	0.27	***	0.42	0.56	-0.14	**	0.21	0.34	-0.13	***	0.39	0.47	-0.08	*	0.00	0.01	-0.01	†
FRL	0.58	1.00	-0.42	***	0.87	0.97	-0.10	***	--	--	--	--	0.73	0.80	-0.07	**	0.72	0.74	-0.02	
IEP	0.00	0.00	0.00		0.08	0.16	-0.08	**	--	--	--	--	0.08	0.04	0.04	*	0.16	0.16	0.00	

Note: For language, all sites collected the PPVT, except NYC which used the ROWPVT. For math, all sites collected the W-J Applied Problems subscale at baseline and follow-up, except NYC which used the ECLS-B math assessment at baseline. For inhibitory control, all sites collected the Pencil tap, except Boston which used Hearts and Flowers. For Seattle, non-English home language was defined by DLL status. FRL=eligible for free-reduced-priced lunch; IEP=has an Individualized Education Plan. HL=home language ***p<0.001 **p<.01 *p<0.05; †p<0.10. Standard deviations for continuous variables are shown in Appendix B, Table 3.

Table 3: Teacher demographics by locality and setting

	Boston			New York City			Seattle			New Jersey			West Virginia		
	Public	CBO	Diff.	Public	CBO	Diff.	Public	CBO	Diff.	Public	CBO	Diff.	Public	CBO	Diff.
Female	1.00	0.80	0.20	0.93	1.00	-0.07 *	0.95	0.96	-0.01	--	--	--	--	--	--
Age (in years)	42.19	36.50	5.69	45.46	38.28	7.19 **	--	--	--	39.23	43.28	-4.05 **	--	--	--
Race/ethnicity															
Asian	0.08	0.10	-0.02	0.04	0.00	0.04	0.05	0.13	-0.08	--	--	--	--	--	--
Black	0.22	0.40	-0.18	0.31	0.50	-0.19	0.05	0.18	-0.13	0.11	0.08	0.03	--	--	--
Latino	0.14	0.10	0.04	0.33	0.40	-0.07	0.00	0.20	-0.20 *	0.16	0.49	-0.33 ***	--	--	--
White	0.47	0.30	0.17	0.27	0.05	0.22 **	0.84	0.40	0.44 ***	0.72	0.35	0.37 ***	--	--	--
Other	0.08	0.10	-0.02	0.05	0.05	0.00	0.05	0.09	-0.04	0.01	0.08	-0.07 **	--	--	--
Education level															
Less than BA	0.10	0.10	0.00	0.00	0.00	0.00	0.00	0.14	-0.14	--	--	--	0.00	0.00	0.00
BA	0.10	0.60	-0.50 *	0.03	0.46	-0.42 **	0.33	0.70	-0.37 **	0.60	0.59	0.01	0.48	0.56	-0.08
MA+	0.80	0.30	0.50 **	0.97	0.54	0.42 **	0.67	0.16	0.51 ***	0.40	0.41	-0.01	0.52	0.44	0.08
Highest degree is in ECE	0.46	0.67	-0.21	0.75	0.63	0.12	--	--	--	--	--	--	--	--	--
Any state teaching cert.	0.98	0.20	0.78 ***	--	--	--	--	--	--	0.99	0.97	0.02	--	--	--
ECE state teaching cert.	0.83	0.20	0.63 ***	0.90	0.48	0.42 **	--	--	--	--	--	--	--	--	--
N yrs teaching experience	15.19	13.31	1.88	18.42	10.50	7.93 ***	--	--	--	--	--	--	13.72	12.79	0.93
N yrs teaching at current school	8.35	8.64	-0.29	12.26	5.64	6.62 **	--	--	--	--	--	--	--	--	--
N yrs teaching PreK	9.05	8.20	0.85	7.61	8.89	-1.28	--	--	--	9.19	10.51	-1.32	--	--	--
Speaks Spanish	0.10	0.00	0.10 *	0.27	0.32	-0.05	0.00	0.13	-0.13	0.14	0.42	-0.28 ***	--	--	--
Speaks a non-Spanish and non-English language	0.24	0.20	0.04	0.16	0.23	-0.07	0.05	0.13	-0.08	0.01	0.08	-0.07 **	--	--	--

Note: SDs for age were, by public school and CBO respectively: Boston (9.37, 8.75); NYC (9.87, 10.11); NJ (9.97, 10.06). SDs for N yrs teaching experience, by public school and CBO respectively: Boston (9.34, 6.68); NYC (8.56, 6.98); WV (8.48, 6.99). SDs for N yrs at current school, by public school and CBO respectively: Boston (7.93, 3.56); NYC (7.63, 7.70). SDs for yrs teaching PreK, by public school and CBO respectively: Boston (7.42, 5.05); NYC (8.29, 7.13); NJ (6.68, 4.86). ***p<0.001 **p<.01 *p<0.05; †p<0.10

Table 4: Classroom structural characteristics and process quality by locality and setting

	Boston				New York City				Seattle				New Jersey				West Virginia			
	Public	CBO	Diff.		Public	CBO	Diff.		Public	CBO	Diff.		Public	CBO	Diff.		Public	CBO	Diff.	
Classroom structural characteristics																				
Class size	15.65	13.45	2.20	†	15.15	16.22	-1.07	*	16.89	16.43	0.46		14.11	14.01	0.1		17.04	17.17	-0.13	
Ratios	2.56	2.8	-0.24	†	5.78	5.13	0.65	†	--	--	--	--	--	--	--		--	--	--	
Classroom quality (CLASS)																				
CO	5.45	5.01	0.44	*	5.62	5.77	-0.16		6.40	6.19	0.21		5.58	5.39	0.19	†	5.17	4.92	0.25	
ES	5.57	5.29	0.28		5.77	6.09	-0.32	†	6.79	6.55	0.24	**	5.95	5.77	0.18		5.74	5.47	0.27	
IS	3.22	3.07	0.15		2.52	2.44	0.07		3.46	3.08	0.38	†	2.52	2.96	-0.44	**	2.74	2.43	0.31	†

Note: CO=Classroom Organization; ES=Emotional Support; IS=Instructional Support. ***p<0.001 **p<.01 *p<0.05; †p<0.10. Standard deviations are in Appendix B, Table 4.

Appendix A: Sampling and data collection procedures details by locality

Sampling

Boston. Our Boston sample consists of 386 students attending the Boston Public Schools (BPS) prekindergarten program or a district-affiliated community-based organization (CBO) preschool program during the 2016-2017 school year. We recruited students from 41 public prekindergarten classrooms and 10 CBO classrooms (1 receiving Head Start funding), nested within 20 public schools and 10 CBO centers during their four-year old year. We randomly selected 25 schools from the 76 schools in the broader district offering the public prekindergarten program; 21 agreed to participate. We used one school as a pilot school for developing new measures and the remaining 20 schools made up the public school sample. We also asked 10 of the 11 district-affiliated CBOs in Boston implementing the BPS prekindergarten model (which was supported by funding from the federal Preschool Development Grant program) to participate in the study and they all agreed. We randomly selected 10 of the CBOs to participate in order to meet our target sample size.

We asked all prekindergarten teachers assigned to general education or inclusion classrooms in each of the 20 public schools to participate in the study in the fall of 2016. We also randomly selected one classroom serving four-year old students within each CBO to participate. Ninety-six percent ($N = 51$) of teachers across public schools ($N = 41$) and CBOs ($N = 10$) agreed to participate in the study activities, including allowing children in their classroom to participate in direct assessments with the research team. After recruiting schools and classrooms, we attempted to collect active consent for all prekindergarten students enrolled in participating classrooms. Research staff met with participating teachers to send home backpack mail providing an overview of the study and a blank consent for the parent to complete and return to

the child's classroom. Field staff then made regular visits to participating classrooms to pick up these consents and document them. Recruitment activities began in late September 2016 and were completed by late November 2016. Eighty-one percent of all children in participating classrooms consented to enroll in the study. Of the total number of children who consented, we randomly selected 50% (~6 – 10 per classroom) to participate in student-level data collection activities for a total sample size of 386 in the fall of 2016.

New York City. The NYC sample consists of 671 students enrolled in CBO and public school sites that were part of a control group as part of the larger Making Prekindergarten Count study (Morris et al., 2016). The research team recruited 69 prekindergarten sites receiving public funding from the city of New York (either from the NYC Department of Education or the Administration for Children's Services) to participate in the study in the spring of 2013. Sites had to be located in a low-income community school district, serve a low-income population of 4-year-old children with significant representation of dual language learners, offer a full-day prekindergarten program, and to have been open for at least two years. The team excluded any programs that reported delivering intensive math curricula. All participating sites received a one-time payment of $5000 to thank them for their involvement. The selected sites reflected the geographical, racial, and ethnic diversity of New York City's low-income population. The final sample included 22 prekindergarten programs in community-based organizations (including some receiving Head Start funding) and 47 public schools spread across four of New York City's five boroughs. The current study only includes the sites randomly assigned to the control group in order to mirror typical business as usual practice as closely as possible. As such, the analytic sample of 34 prekindergarten sites includes 23 public schools and 11 community-based organizations.

After recruiting sites, the research team asked all prekindergarten lead teachers and assistant teachers within each site to participate in the study. Eligible classrooms had to be general education/inclusion serving only four-year old students. The final classroom sample size included in the current study was 86, including 61 public schools and 25 CBO classrooms. Classrooms participated in the study for two academic years, in 2013 – 2014 and 2014 – 2015. The current study reports on year two observational data from classrooms collected during the winter and early spring of 2015 and teacher survey data collected in the spring of 2015.

In the fall of 2014, the research team attempted to collect active consent for all prekindergarten students enrolled in participating classrooms. Research staff sent home backpack mail providing an overview of the study and a consent form for parents to complete and return and then visited regularly to pick up completed consent forms. Ninety-five percent of children in participating classrooms consented. The team randomly selected about half of these children (N = 691) to participate in assessments in the spring of the prekindergarten year and a subset of those students (N = 334) to participate in assessments in the fall of the prekindergarten year. Students in the study sample were representative of the broader population of consented students and the students enrolled in the participating sites. However, students enrolled in the broader NYC UPK program were more diverse than students in the current study. During the 2014 – 2015 year, there were 52,741 children enrolled in the city's UPK program. Of these students, 37% were Hispanic, 30% were Black, 17% were White, 13% were Asian, and 3% were Native American or multi-racial (Potter, 2016).

Seattle. The Seattle sample consists of 1,142 3- and 4-year-old students who attended the Seattle Preschool Program (SPP) during the 2018-2019 school year and participated in the fourth year of an evaluation of SPP (364 three year olds (32%) and 778 four year olds (778)).

Participants were invited to participate by the City as part of the registration process. All 11 schools and 38 CBOs offering SPP were included in the sample. There were a total of 75 classrooms that participated in the study, including 19 in public schools and 56 in CBOs. To maintain a consistent coding scheme for Head Start locations, five classrooms in four locations affiliated with Seattle Public Schools were coded as CBOs. Consent forms were distributed to all children enrolled in SPP and 99% consented to participate in the study. Twelve students per classroom were randomly selected to be included in the study, from among those who consented. A total of 306 children were in public schools and 836 in CBOs. Another 47 children were enrolled in Family Child Care and are not included in these analyses.

New Jersey. The NJ sample consists of 1,116 3- and 4-year-old students enrolled in state-funded preschool in 15 districts voluntarily participating in a larger study about implementing and sustaining high quality pre-K (454 three year olds (41%) and 659 four year olds (59%)). Districts participated in the study over two cohorts: cohort one during the 2017-2018 school year and cohort 2 in 2018-2019. Districts in the sample were selected to ensure variation in district characteristics, geography, experience providing high-quality preschool, and third grade test scores. Six districts had been implementing high-quality state-funded preschool since it was mandated by the Abbott vs. Burke court decisions in 1998. The other nine districts began operating the state's high-quality preschool model more recently either through federal Preschool Development Grants or the state's expansion funding. New Jersey's state-funded preschool programs use a mixed-delivery model in which school districts can subcontract with Head Start and private child care providers to provide state-funded preschool. Districts make their own decisions on the degree to which to partner with community providers, and which providers they partner with. Across districts, 47 schools and 38 CBOs were randomly selected to be included in the sample. In small districts, all schools and CBOs with preschool were included

in the sample. Classrooms were then randomly selected to be included in the sample. In small districts and small schools/CBOs, all preschool classrooms were included in the sample. A total of 277 classrooms (180 in schools and 97 in CBOs) were included in the sample. Of the CBOs, 39 were Head Start programs.

Our research team attempted to collect active consent for all students enrolled in the selected classrooms through meetings with district ECE staff, teachers, and communication with parents, including sending home backpack mail with information on the study and consent forms. Research staff fielded questions from parents about the study. An average of four children in each classroom from among those whose parents consented were randomly selected to participate, resulting in a sample with 732 children in schools and 387 children in CBOs.

West Virginia. The West Virginia sample consists of 606 4-year-old children enrolled in West Virginia's Universal Pre-K programs who participated in a regression discontinuity and longitudinal evaluation of the program. Children in the study attended pre-K in 2015-2016 in seven counties that had lower rates of enrollment in pre-K during the prior year (so that a comparison group was available for the longitudinal evaluation). West Virginia's state-funded preschool programs use a mixed-delivery model in which counties can subcontract with Head Start and private child care providers to provide state-funded preschool. Counties make their own decisions on the degree to which to partner with community providers, and which providers they partner with. Early in fall 2015, schools (N = 45 public school sites and 16 CBOs; 1 CBO was a Head Start program) in the 7 counties were informed about the study and families were invited to participate. In five of these counties, all preschool programs and classrooms were included, and in the two larger ones, a random sample of programs were included. In each school, consent forms were sent home to all students in all pre-k classrooms to enroll children for participation through active consent. An average of 6.3 children per pre-K classroom participated in the study

(ranging from 1 to 16 children). There were 401 pre-K students attending in the public schools and 205 in CBOs. In the study year, if 40% or more of the children in a school were certified as eligible for free/reduced-price meals, then all children in the school were categorized as “low SES” (a policy that later changed in the 2017-2018 school year).

Procedures

Boston

Direct assessments. Prekindergarten students were assessed in the fall of 2016 (October 1st through December 12th) and the spring of 2017 (April 5th through June 16th). We assessed kindergarten children in the fall of 2017 (October 1st through December 12th) and spring of 2018 (April 5th through June 16th). All child assessors were trained to reliability. A master's-level supervisor observed 10% of field assessments to ensure high-quality administration.

Before beginning the study battery, in both prekindergarten and kindergarten, assessors used the Pre-language Assessment Scale (preLAS; Duncan & DeAvila, 1998) to determine the administration language for a subset of assessments. Of the 378 children in the prekindergarten study sample, 43 (11%) completed a subset of assessments in Spanish in fall 2016, and 15 (4%) completed assessments in Spanish in spring 2017. There were N = 369 children who completed the assessments in fall, N = 357 who completed assessments in spring, and N = 348 who completed assessments at both time points. See more in McCormick et al. (2021).

Classroom observations. During the winter of 2017, each classroom was videotaped for two hours during two visits. Visits were scheduled in advance with teachers. Observers participated in a two-day training to learn the CLASS measure and then established reliability on a set of master codes created by the test developers. Coders started coding the tapes once instructional time began. As recommended by the measure's protocol (Pianta et al., 2008), coders used cycles of 20 minutes for observing and 10 minutes for scoring, which they repeated 4 times for each observation. Scores across the four segments were first averaged to calculate observation-specific scores and then the scores across observations were averaged to generate

one overall score for each classroom. The team double-coded 20% of the observations to assess interrater reliability. The final ICCs representing interrater reliability were 96% for Emotional Support, 94% for Classroom Organization, and 88% for Instructional Support. We also did a drift check wherein coders had to code a master tape every three weeks to ensure they were still reliable before continuing to codetapes.

Teacher surveys. In the spring of the PreK year, we asked teachers to complete a survey reporting on their demographic characteristics, teaching experience, and instruction.

Parent surveys. We collected parental demographic information via parent surveys in the fall of prekindergarten (fall of 2016). We used email and text message to contact the parents of the consented study students to ask them to complete the 20-minute parent survey. We also sent biweekly reminders to the parents asking them to complete the survey and used backpack mail for parents who did not complete the survey electronically. Although the majority of parents completed the survey in English, we also translated the survey into Spanish, Vietnamese, and Mandarin. Parents received a $25 gift card for completing the survey. Of the 378 students in the first year of our study, 355 (94%) had parents who completed the survey and had some valid parent-reported data used to create key covariates.

Administrative data. We used administrative data to create child-level variables.

New York City

Child assessments. The research team trained a team of data collection staff to reliability on direct assessments of children's math, language, executive functioning, and self-regulation skills. Data collectors participated in a five-day training on these assessments, engaged in

multiple practice assessments with trainers, and then had to pass two mock assessments – one with another adult and a second with a four- or five-year old child recruited to participate in training activities. The team certified data collectors if they administered both assessments with at least 90% reliability, based on an administration quality checklist. These field staff then administered direct assessments in the fall and spring of PreK and the spring of kindergarten. The research team used the Pre-Language Assessment Scale (preLAS) (Duncan & DeAvila, 1998) as a warm-up to the assessment battery and to determine the administration language for a subset of assessments during the PreK year only (Barrueco, Lopez, Ong, & Lozano, 2012). Of the children in the current study sample who completed assessments – making up just the control group in the larger impact study – 18% did not pass the preLAS and completed all assessments in Spanish during the fall of PreK. Eight percent of students completed all assessments in Spanish in the spring of PreK. There were no Spanish assessments in the spring of kindergarten. Children received a book to thank them for participating in the assessments.

Classroom observations. Observations took place in a subset of classrooms ($N = 69$ at least one classroom randomly selected per site) in the spring of 2013, prior to the beginning of the intervention, in order to establish baseline equivalence. All classrooms participated in observations in the spring of the PreK year to measure classroom outcomes. Prior to the start of each data collection wave, the research team trained observers to reliability on the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) and a version of the Classroom Observation of Early Mathematics – Environment and Teaching (COEMET) adapted with the measure's developers for this study (Morris et al., 2016).

The team worked with the developer of the CLASS to certify observers. They completed a two-day training with a certified master trainer. Observers then rated a series of master-coded

video clips on the dimensions of the CLASS. They needed to demonstrate 80% agreement "within 1" across the ten dimensions in order to be considered reliable and able to collect live CLASS data in the field. For the adapted version of the COEMET, the team conducted a five-day training for observers where they completed practice coding activities using master-coded videotapes. Coders needed to agree with 80% of the master codes "within 1" for quality items rated on Likert scales and demonstrate 80% exact agreement across videos on the number of total math activities and the practices done in the math activities. Finally, coders needed to agree with the overall amount of time spent in math activities with no more than 20% difference with the master coder. Coders then completed these same reliability exercises during a live observation with a master coder before collecting data in the field. Observers then conducted two separate 3-hour observations in each PreK classroom during morning instruction. They coded the CLASS measure during the first observation and the adapted version of the COEMET during the second.

Parent survey. Parents provided information on their and their child's demographic characteristics in the fall of 2014. Research staff sent home backpack mail providing an overview of the study, a consent form, and a short questionnaire to capture parent-reported demographic and household information. Research team members then visited the school regularly to pick up completed forms and were able to collect them for 95% of children enrolled in participating sites.

Teacher data. When consenting to participate in the study, teachers reported on their demographic characteristics, teaching credentials, stress, and teaching experience. They received a $20 gift card to thank them for their time completing the survey.

New Jersey

Child assessments. Children were assessed in the fall between September and November and again in the spring between May and June. Spanish-speaking children were assessed in English and in Spanish on the PPVT and Woodcock Johnson (by Spanish-speaking bilingual data collectors). Teachers were the primary informants of whether or not a child should be assessed in Spanish in addition to English. However, if a Spanish bilingual child did not get past set 3 in the PPVT in English, they were also assessed in Spanish. The Pencil Tap was administered only once, in the child's dominant language. All data collectors were trained by NIEER staff during a two-day training in the fall prior to the start of data collection. Each assessor was shadow scored to ensure 100% accuracy in assessment. A refresher training took place prior to the beginning of spring data collection.

Classroom observations. Each classroom was observed on the CLASS between February and March, by a team trained to reliability standards. Spanish bilingual data collectors conducted observations in classrooms with Spanish-speaking teachers and children. Classroom observations occurred during the first three hours of the school day. CLASS reliability was done through video coding per protocols from the instrument developer. Observers were deemed reliable when they reached at least 85% exact agreement or within one with the master coder. This was followed by two cycles of in person reliability at the beginning of data collection, and with online video calibrations mid-point through data collection.

Child demographics. Parent surveys were initially distributed in the fall and could be completed either online or on paper in English and Spanish. Parents received a $20 gift card for completing the survey. NIEER staff followed-up with parents and worked with schools

throughout the school year to obtain the survey information. A short version was sent to attempt to collect some basic information from parents who had not completed the survey. In addition, some demographic data were provided by the school districts (administrative data).

Teacher survey. Lead classroom teachers completed a survey during the spring in which they answered questions about their background, experience teaching, highest degree, and language spoken.

Seattle

Child assessments. Children enrolled in the Seattle Preschool Program were assessed by trained data collectors in fall 2018 and again in spring 2019, with a minimum of six months between the two assessments. All child assessors were trained to reliability by the research team. Data collectors were trained during a two-day training, were given several days to practice, and were tested for reliability prior to beginning data collection in the fall and in the spring. Reliability was checked again half-way through each data collection period. All assessments were completed in English with the exception ten Spanish-speaking children were also assessed on the battery in Spanish. Classroom teachers were the main informants regarding children's languages spoken.

Classroom observations. Direct observations of classroom practices and processes were conducted. between February 2019 and April 2019 by trained and reliable observers. CLASS observers were trained by a CLASS certified trainer and all met the Teachstone reliability certification requirements. Observers' reliabilities ranged from 92 to 98%.

Child demographics. Information on child characteristics was collected through administrative data . Additionally, teachers were the primary informant on children's home language.

Teacher characteristics. Information on teachers' race/ethnicity, languages spoken, and gender was obtained from administrative data. Information on teachers' highest degree was collected via a brief teacher survey conducted at the time of classroom observations.

West Virginia

Child assessments. Children were assessed in Fall 2015 when enrolled in the West Virginia Universal Pre-K program and again in Fall 2016 as kindergarteners. NIEER and Marshall University worked collaboratively to hire and train child data collectors. Data collectors were trained on all child assessments over a two-day training. After completion of the training, data collectors were shadowed twice while administering the child assessments by expert staff. All data collectors then obtained 100% reliability. All children were tested in English as almost none of the children were dual language learners (see Table 2 in the main text).

Classroom observations. Preschool classroom observations occurred between February and May 2016 by reliable observers. Observers were trained in a full-day training. CLASS observers were trained by a CLASS certified trainer and met the Teachstone reliability requirements with agreement percentages ranging from 84 to 100%. Observers were required to go through online video calibration mid-point through data collection.

Child demographics. Child demographics information was collected through the West Virginia Education Information System (WVEIS). In the study year, if 40% or more of the

children in a school were certified as eligible for free/reduced-price meals, then all children in the school were categorized as “low SES”(a policy that later changed in the 2017-2018 school year).

Teacher survey. A brief survey was administered to teachers at the time of classroom observations, which included information on teacher highest degree and teaching experience.

Appendix B: Additional measures details

CLASS. The CLASS and these three constructs – Emotional Support, Classroom Organization, and Instructional Support – generally show good psychometric validity (though the CLASS is not consistently related to gains in children's outcomes in the prekindergarten year (e.g., Guerrero et al., 2019).

Vocabulary. The PPVT-IV has been normed and used widely in diverse samples of children in the U.S (Puma, Bell, Cook, Heid, & U.S Department of Health and Human Services, 2010), and it has shown qualitative and quantitative validity properties (Dunn & Dunn, 2007). The test–retest reliability ranges from 0.92 to 0.96. It requires children to choose (verbally or nonverbally) which of four pictures best represents a stimulus word. The ROWPVT- 4 is a nationally normed measure that has been used widely in diverse samples of young children and has been shown to be reliable and valid (Martin & Brownell, 2011).

Math. The Woodcock–Johnson Applied Problems III (Woodcock et al., 2001) subtest requires children to perform relatively simple calculations to analyze and solve arithmetic problems. Its estimated test–retest reliability for 2- to 7-year-old children is 0.90 and it has been nationally normed and used with diverse populations of children. The *ECLS-B* math assessment is a 44-item direct assessment that measures a range of early math skills including number sense, number properties, and operations; measurement; geometry and spatial sense; and patterns, algebra, and functions. It has been used in large-scale studies with diverse populations including the ECLS-B and the Head Start FACES study and shown to demonstrate good psychometric properties (reliability of 0.89 in Pre-K; Najarian, Snow, Lennon & Kinsey, 2010).

Executive function. The Pencil Tap assessment requires children to tap once immediately after the assessor taps twice and vice versa. It has demonstrated convergent and

predictive validity as well as test–retest reliability in prekindergarten samples (e.g., r = 0.80; Lipsey et al., 2017). It has also been widely used with diverse samples of young children in prekindergarten programs (e.g., Goldman et al., 2019; Weiland & Yoshikawa, 2013). The Hearts and Flowers test combines the cognitive demand of the Simon Says (Duncan & De Avila, 1998) and spatial Stroop tasks (Hilbert, Nakagawa, Bindl, & Bühner, 2014). During congruent – or hearts – trials, children need to obey the rule: "Press on the same side as the stimulus" and during incongruent – or flowers – trials, children follow the opposite rule: "Press on the side opposite the stimulus" (Wright & Diamond, 2014).

Appendix C: Additional tables and figures

Figure 1: Map of mixed-delivery state-funded public prekindergarten systems in the U.S. (based on the authors' analysis of NIEER data from 2020-2021; Friedman-Kraus et al., 2022)

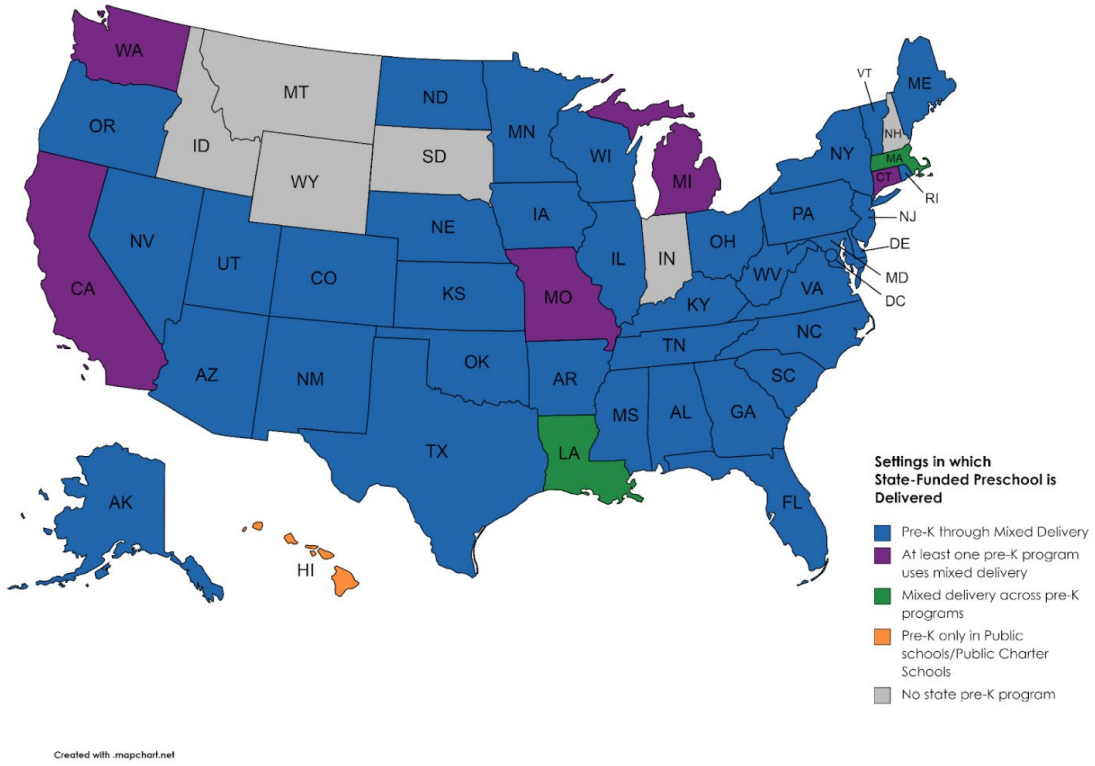
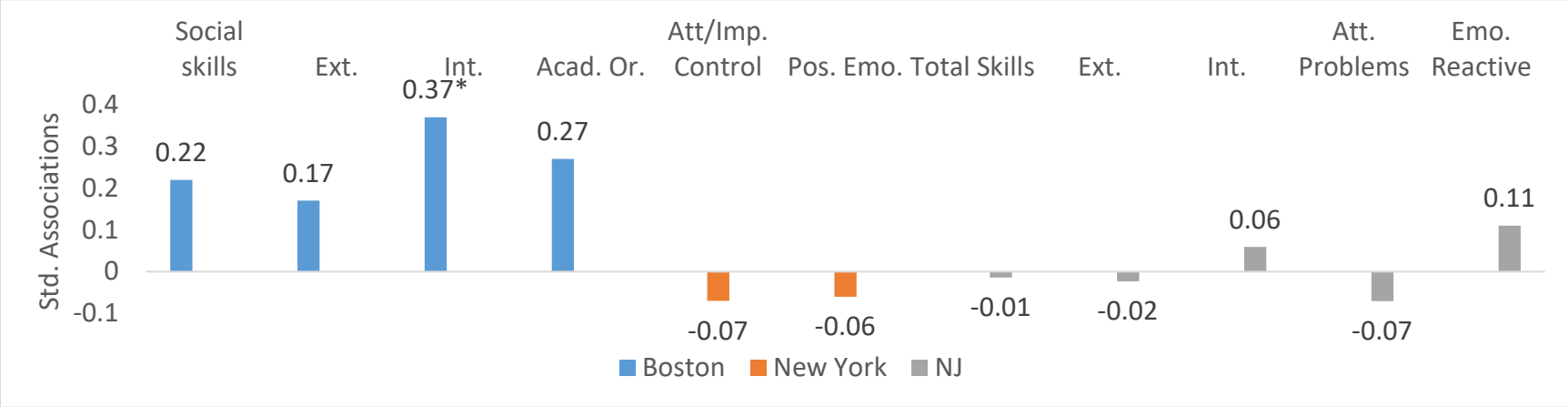


Figure 2: Standardized associations between attending a CBO site (versus public school) and children's social-emotional skill gains in public prekindergarten in Boston, NYC, and NJ



Note: Higher scores on externalizing and internalizing behaviors indicate more of these behaviors (i.e., negative valance). Ext.=Externalizing; Int.=Internalizing; Acad. Or.=Academic Orientation; Att/Imp. Control=Attention/Impulse Control; Pos. Emo=Positive Emotion; Att. Problems=Attention Problems; Emo. Reactive=Emotionally Reactive. In Boston, we used teacher-reported data from three subscales of the Social Skills Inventory System (SSIS; Gresham, Elliott, Vance, & Cook, 2011) and from the Task Engagement Subscale of the Teacher-Child Rating Scaling (Hightower et al., 1987). In New York City, after assessors finished testing children, they also rated students' behavior on the Attention/Impulse Control and Positive Emotion subscales of the Preschool Self-Regulation Assessment (Smith-Donald et al., 2007). In NJ, teachers completed the Caregiver Teacher Report Form (CTRF; Nores et al., 2022).

Table 1: Associations between attending a CBO site (versus public school) and children's gains in public prekindergarten

	Boston			New York City			Seattle			New Jersey			West Virginia		
	CBO	SE	*P*-value	CBO	SE	*P*-value	CBO	SE	*P*-value	CBO	SE	*P*-value	CBO	SE	*P*-value
Vocabulary	1.17	2.63	0.66	-3.64	1.73	0.04	0.47	1.54	0.76	1.34	1.69	0.43	-3.38	1.79	0.06
Math	-1.27	0.55	0.03	-0.18	0.34	0.59	-0.27	0.32	0.41	0.30	0.28	0.28	-0.17	0.34	0.60
Inhibitory control	-0.04	0.03	0.23	0.01	0.03	0.59	-0.01	0.02	0.52	-0.01	0.04	0.69	-0.02	0.02	0.22

Note: NJ and WV models included district fixed intercepts.

Table 2: Robustness of NJ and WV results to alternative modeling decisions

	NJ - Classroom RI			NJ - District RI			WV- District FI			WV - District RI		
	CBO	SE	*P*-value	CBO	SE	*P*-value	CBO	SE	*P*-value	CBO	SE	*P*-value
Vocabulary	-0.14	1.20	0.91	0.71	1.31	0.59	-3.43	1.81	0.06	-3.38	1.45	0.02
Math	0.09	0.22	0.69	0.16	0.25	0.53	-0.21	0.34	0.53	-0.22	0.29	0.46
Inhibitory control	-0.03	0.02	0.15	-0.03	0.02	0.15	-0.02	0.02	0.18	-0.02	0.02	0.12

Notes: FI=Fixed Intercepts; RI=Random intercepts. NJ District Fixed Intercepts and WV District Fixed Intercepts models are the primary results shown in Figure 1 of the main text and Appendix C Table 1.

Table 3: Standard deviations for continuous child baseline, outcome, and demographics by locality and setting

	Boston		NYC		Seattle		NJ		WV	
	Public	CBO	Public	CBO	Public	CBO	Public	CBO	Public	CBO
Panel 1: Child skills										
Language										
Baseline	28.67	21.57	19.84	18.92	27.27	27.21	23.97	24.55	19.15	19.97
Follow-up	26.49	23.05	18.02	20.45	26.79	27.20	23.45	24.68	18.34	18.54
Math										
Baseline	5.18	4.53	6.60	6.27	5.36	5.33	4.08	4.15	3.87	4.31
Follow-up	4.52	4.91	4.43	4.27	4.99	5.05	4.21	4.31	3.51	3.67
Inhibitory Control										
Baseline	0.18	0.14	0.35	0.34	0.35	0.34	0.33	0.32	0.23	0.24
Follow-up	0.21	0.17	0.31	0.30	0.35	0.35	0.38	0.36	0.17	0.19
Panel 2: Demographics										
Age (in years)	0.29	0.34	0.29	0.31	0.49	0.54	0.54	0.57	0.43	0.38

Note: For language, all sites collected the PPVT, except NYC which used the ROWPVT. For math, all sites collected the W-J Applied Problems subscale. For inhibitory control, all sites collected the Pencil tap, except Boston which used Hearts and Flowers.

Table 4: Standard deviations for classroom structural characteristics and quality by locality and setting

	Boston		NYC		Seattle		NJ		WV	
	Public	CBO	Public	CBO	Public	CBO	Public	CBO	Public	CBO
Classroom structural characteristics										
Class size	3.87	2.30	1.92	2.96	2.02	3.67	1.6	4.15	2.78	2.61
Ratios	0.70	0.26	1.80	1.72	--	--	--	--	--	--
Classroom quality (CLASS)										
CO	0.59	0.43	0.76	0.69	0.53	0.53	0.72	0.68	1.18	1.13
ES	0.60	0.51	0.82	0.76	0.36	0.29	0.69	0.68	0.90	0.89
IS	0.63	0.65	0.80	0.77	0.87	0.83	0.86	0.98	0.85	0.75

Note: CO=Classroom Organization; ES=Emotional Support; IS=Instructional Support.