Blueprint Labs

Discussion Paper #2023.20

Lottery evidence on the impact of preschool in the United States: A review and meta-analysis

Jesse Bruhn
Emily Emick

November 2023

MIT Department of Economics
77 Massachusetts Avenue, Bldg. E53-390
Cambridge, MA 02139

National Bureau of Economic Research
1050 Massachusetts Avenue, 3rd Floor
Cambridge, MA 02138

Lottery evidence on the impact of preschool in the United States: A review and meta-analysis*

Jesse Bruhn and Emily Emick

November 2023

Abstract

This paper reviews research on the design and effectiveness of preschools in the United States. Three randomized controlled trials that enrolled roughly 350 total children in demonstration studies during the 1960s and 1970s shape much of the current discussion of preschool. We examine what these studies and more recent ones based on random assignment research designs reveal about preschools today.

We find that the broad conclusions of the demonstration program literature are remarkably robust. Preschool generates large initial improvements in academic outcomes that fade out over time. The beneficial effects of preschool then re-emerge later in life on academic outcomes, like high school graduation, and non-academic outcomes, like criminal justice contact. Modern studies using larger samples also find positive effects on school discipline, SAT taking, and college attendance.

Variations in magnitude between the demonstration programs and modern studies can be explained by (1) differences in the quality and intensity of the interventions; and (2) the expansion of preschool, which changes the nature of the counterfactual for students not in preschool. Both factors highlight the importance of quantifying the features that create high-quality preschool experiences. Scaling effective practices have the potential to benefit the large number of current preschoolers and new children enrolled as a result of expansion.

Despite the concordance of findings between the demonstration and modern-era programs, the overall volume of evidence on preschool based on random assignment remains thin. Over the last 60 years, only seven experiments have randomized access to preschool. Since 1990, a total of 14 studies have randomly varied preschool characteristics. By leveraging lotteries used to allocate seats in oversubscribed schools, researchers can evaluate the impact of preschool in a broader variety of settings as well as identify the drivers of quality.

*Jesse Bruhn: Department of Economics, Brown University. Email: jesse_bruhn@brown.edu. Emily Emick: Department of Economics, Brown University. Email: emily_emick@brown.edu.

Preschool enrollment in the United States has grown dramatically over the last 70 years. Starting from negligible levels in 1950, as of 2018, 61 percent of 4-year-olds and 34 percent of 3-year-olds attended some form of center-based preschool (Cascio, 2021). Much of this surge occurred in conjunction with state and federal expansion of seats (Cascio, 2021). Head Start now enrolls 7 percent of all 4-year-old children at a cost of more than 10 billion USD. As of 2019, 44 states had some form of large-scale public preschool program (US Department of Health and Human Services, n.d.; Friedman-Krauss et al., 2019).

The increase in public support and contemporaneous rise in enrollment has undoubtedly been driven by a belief among the American public that early childhood education holds the potential to generate an enormous return on investment. As President Barack Obama put it in 2017: "[For] every dollar we put into high-quality early childhood education we get $7 back in reduced teen pregnancy, improved graduation rates, improved performance in school, reduced incarceration rates," (Obama, 2015). Still, nearly half of all children ages 3 and 4 *are not* currently enrolled in preschool (De Brey et al., 2021). State legislatures nationwide are expanding preschool: A recent analysis found that 14 states are currently discussing preschool expansion, with 11 likely to pass some form of universal eligibility within the next calendar year (Potts, 2023). The impact of this wave of legislation will depend on the features of the individual programs and the causal effects they generate on the children served.

What do we know about the causal effect of preschool in the United States? Some evidence cited in support of claims like the one made by President Obama comes from non-experimental studies (e.g., Reynolds et al., 2011). However, observational studies may be biased by selection. On the other hand, much of the commonly cited evidence based on high-quality, lottery variation relies on the results of three randomized controlled trials that were conducted nearly 50 years ago and that enrolled a total of just over 350 children from families intentionally selected to be nonrepresentative of the population as a whole.[1] The approach to statistical inference in the demonstration program literature has also been the subject of criticism and controversy (Anderson, 2008; Conti et al., 2016; Heckman et al., 2010b). Thus, it is unclear how much weight policymakers today should place on the findings from these demonstration studies.

In this paper, we review lottery-based evidence about preschool in the United States. In contrast to existing reviews and meta-analyses of the early childhood education literature, we forgo a broad approach incorporating multiple research designs and populations

[1]See Anderson (2008) for a concise summary of all three RCTs.

of study. Instead, we take a narrow but deep dive into the body of work based on random assignment. This allows us to compare results from the small-scale and older demonstration studies with current work using similar methods and larger samples. This "gold standard" evidence can form a solid foundation for policymaking and serve as a springboard for future research.

Papers needed to meet five criteria to be included in this review. First, the paper had to use lottery variation to evaluate early childhood education in the United States. Second, the intervention studied had to have components that target childhood development during the typical preschool age ranges (usually 3-5). Third, the study had to examine outcomes that pertained directly to children, as opposed to impacts on parents or others. Fourth, the paper must have been published in a peer-reviewed journal after 1990. Fifth, the paper had to report the data necessary to calculate effect sizes. We searched for papers that met these criteria using a modified version of the discovery protocol outlined in Jackson and Mackevicius (2023). Ultimately, we found 44 papers describing results from 21 distinct experiments that fully met the criteria.

We begin, in section 2, with a novel re-analysis of key results from the demonstration program literature. To address concerns related to small sample sizes, sub-group analysis, and multiple testing, we pool the published estimates across all sub-groups and across all three interventions: Perry Preschool Project, the Early Training Project, and Carolina Abecedarian Project. Consistent with prior work, we find large impacts in test scores at age 5 that fade in magnitude over time. However, unlike the prior analyses, we find statistically precise evidence from the pooled averages that demonstration programs generated impacts on IQ scores that persist into high school. The pooled averages also provide statistically precise evidence that the programs boosted high school graduation, reduced teen pregnancy, and reduced the use of illegal drugs. While the pooled averages cannot detect statistically precise evidence of effects on other long-run outcomes, we do find evidence of an overall improvement in an index of adult well-being. In light of these findings, we also discuss recent evidence related to the intergenerational effects of Perry Preschool, as well as prominent theories regarding the role of non-cognitive skills (e.g., focus, curiosity, or grit/resilience) in mediating the impact of the demonstration programs on long-run outcomes (García et al., 2023; Heckman et al., 2013). We conclude our discussion of the demonstration programs by noting that the time period of the interventions; their unusual intensity and quality;[2] and the non-representative nature of the study sub-

[2]The demonstration programs had unusually small class-sizes and also involved non-schooling compo-

jects fundamentally limit the direct applicability of this body of work to the questions of interest for early childhood education today. Thus, further study using modern data, interventions, and populations is essential.

Next, in section 3, we discuss results from four modern experiments that randomize offers of admission to oversubscribed preschool programs. The results from this body of work are remarkably similar to many of the findings from the demonstration program literature. Preschool in the modern period causes a dramatic rise in test scores that fades over time, only to re-emerge on consequential later life outcomes such as high school graduation and college attendance. However, the magnitudes of the effects appear to be smaller than those found in the demonstration program studies. We discuss differences in the intensity and nature of the intervention, as well as the changing prevalence of preschool in the control group, as likely explanations. We also point out apparent (and sometimes contradictory) effects on outcomes, like disciplinary violations and juvenile incarceration, that are thought to be a function of non-cognitive skills. We end this section with a discussion of the enormous amount of heterogeneity in site-level impacts found in the literature. This heterogeneity highlights the potential scope for quality improvements to benefit both current preschool students *and* future children drawn into the system by expansion. We believe that uncovering the causal sources of heterogeneity in site-level effects is among the key questions relevant for policy design today.

Yet, despite the importance of understanding the drivers of quality, lottery evidence on "what works" in preschool remains thin. Our search uncovered 14 experiments that randomized preschool characteristics such as professional development, curriculum, hours of instruction, class size, and language immersion. Unfortunately, important caveats attach to much of this literature. Specifically, we find that most of the conclusions in these studies (1) are based on small effective sample sizes; (2) do not test for sample balance; and (3) do not account for heteroscedasticity, clustering, or other forms of dependence across observations in their approach to statistical inference. These concerns raise serious doubts about the statistical reliability of the findings from this body of work. However, with those caveats in mind, we find exciting results from a study that examines the effect of switching from half-day to full-day preschool that we believe merits close attention in future work, as do some of the curriculum-based interventions considered.

We conclude, in section 5, with a discussion of key gaps in the literature and opportu-

nents such as supplemental nutrition, medical care, and parental home visits. See discussion in section 2.1.

nities for future work. Priorities for future lottery studies should be to (1) better understand the practices that can improve the quality of preschool within the existing system; (2) develop novel methods of identifying specific drivers of the outcomes of interest; and (3) explore the overall impact of preschool across a broader range of outcomes in different geographic and sub-group populations. The current wave of state-level legislation offers a generational opportunity for the research community to provide policymakers with an innovative body of credible and high-impact research.

1 Why lottery studies?

In this section, we detail the methods we used to build the body of papers that form the core focus of this review. We begin by reviewing the benefits and limitations of lottery-based research designs. We conclude with a detailed justification of our inclusion criteria and a discussion of the search protocol that generated the corpus of lottery-based papers included in this review.

1.1 The benefits of random assignment

At its core, making appropriate decisions about the size, scope, and design of the early childhood education system in the United States requires *counterfactual reasoning*: What would happen to the standardized test scores of children if the nation had universal preschool as opposed to the status quo? Would moving from half-day to full-day preschool lead to improved readiness for school entry and subsequent performance in the K-12 system? Does access to high-quality preschool experiences generate meaningful improvements in adult health and well-being?

The overarching goal of this paper is to establish a baseline of facts about the causal effects of preschool in the United States. We hope these facts will be used to guide decisions and serve as a springboard for future research. However, establishing this baseline is not straightforward. Naive comparisons between the outcomes of children who do and do not spend time in center-based, early-childhood education are unlikely to provide credible answers due to selection bias. For example, many students attend preschool in programs like Head Start, which are means-tested and thus disproportionately enroll low-income populations. Therefore, a comparison between children who attend preschool and children who do not could be misleading, since the types of families who access

these experiences will be fundamentally different both on observable dimensions like income and on other dimensions that are harder to measure such as parental engagement. In practice, selection bias is the key concern for applied work that seeks to quantify the impact of preschool.

Randomized controlled trials and other forms of lottery-based variation represent an attractive solution to the selection bias problem. Because the offer to attend preschool is assigned randomly, the individuals who obtain access should be otherwise similar to those who apply but do not receive an offer, ensuring an "apples-to-apples" comparison. This is why random assignment is often referred to as the "gold standard" for evidence of causal effects, both in the literature on early childhood education and across the social sciences more broadly (Angrist and Pischke, 2009; Cascio, 2021). Importantly, estimates from random assignment are credible when the lottery occurs naturally, as is often the case in the event of oversubscription or if the random assignment occurs as part of a formal, randomized controlled trial analogous to those deployed in medical research.

Lottery-based estimates of causal effects are not without drawbacks. The first relates to feasibility. Sometimes, it is not possible to randomize certain interventions at scale because of financial constraints. While randomizing the existence of Head Start centers across a large number of regions could address questions regarding the broader, community-wide impacts of preschool, doing so at scale would be prohibitively expensive. Other times, it may not be ethical to randomize access to certain types of interventions and services. For example, safety requirements necessary for licensing undoubtedly place a large financial cost on providers; however, randomizing this feature could jeopardize the safety of vulnerable children.

The second drawback relates to the populations being studied. While estimates based on lottery variation generate compelling causal inferences and hence contain a high degree of *internal validity*, a treatment effect is not always estimated in a time, place, and population that will translate into the domain of a policy decision of interest. As a result, lottery-based evidence may reduce the *external validity* of the conclusions drawn because they pertain to a small slice of the population that was exposed to (and complied with) random assignment.

1.2 Inclusion criteria

To be included as a focus of this review, we decided that a research study would need to meet the following five criteria:

- Use a research design that is based on random assignment to evaluate early childhood education in the United States.
- Have intervention components that target childhood development during the typical preschool-age ranges (usually 3-5).
- Examine outcomes that pertain directly to students.
- Have been published in a peer-reviewed journal sometime after 1990.
- Report the data necessary to calculate effect sizes.

The rationale for the last three criteria is straightforward: We restrict to studies that report student outcomes because we are interested in the impact of preschool on children and not, for example, in how interventions change teacher behavior within the classroom. Publication in a peer-reviewed journal helps to ensure a minimum level of scientific objectivity. Reporting data necessary to calculate effect sizes allows us to compare (and sometimes pool) treatment effects estimated across different outcomes, measurement scales, and study populations.

However, adopting criteria 1 and 2 restricts the scope of the evidence we consider in a more consequential way. For that reason, we will now explain why we included them and discuss how the benefits of doing so outweigh the associated costs.

Restriction to lottery-based research in the United States. The overarching goal of this paper is to establish a baseline of facts about the causal effect of preschool in the United States. However, we acknowledge that an enormous amount of high-quality and credible research exists on the causal effects of preschool that use alternative quasi-experimental and structural methods. We elected not to include these papers as a primary focus of this review for two reasons.

First, some recent review papers and meta-analyses have adopted a broad methodological perspective on the state of evidence in the early childhood education literature (Cascio, 2021; Duncan et al., 2022; Duncan and Magnuson, 2013; McCoy et al., 2017). Consequently, a deep dive into the findings from lottery evaluations adds more value to the existing body of knowledge than a broad survey.

Second, there are clear conceptual benefits to focusing on random assignment. This is an easy-to-define criterion for elevating a finding into the "gold-standard" territory of evidence that we are striving to capture with this review. Once we move away from

random assignment, it becomes much less clear where to draw the line, since different social scientists have different standards for what they view as credible among the alternative approaches. Further, even with a given approach or research design, there is often disagreement about what constitutes sufficiently credible evidence. Restricting this analysis to studies based only on random assignment allows us to avoid making subjective judgment about how many periods of "pre-trends" should be necessary to warrant inclusion, or whether a well-done event study is more or less credible than a weakly powered regression discontinuity design. Thus, by setting the bar to studies that are based on lotteries or other forms of random assignment, it is easier and more transparent to maintain a high degree of objectivity.

Finally, a similar rationale motivates our restriction to the United States. While it is certainly a useful exercise to explore foreign literature and extrapolate from these settings to the United States, including such studies would raise important questions about external validity (see section 1.1). As with the decision to focus on studies that use random assignment, we draw a clear line to improve the objectivity of the analysis and avoid judgment calls regarding which settings and populations are "similar enough" to the United States to warrant inclusion.

Restriction to interventions that target the preschool age group. Social scientists often define early childhood education as group instruction for children younger than the standard age for K-12 enrollment (Cascio, 2021). We opted to narrow the scope of this definition to keep the review focused and afford space to dive deep. In addition, publicly funded preschool for 3- and 4-year-olds is expanding nationwide. For example, a recent news analysis concluded that 14 states were currently discussing preschool expansion, with 7 likely to pass some form of universal eligibility for preschool within the next year (Potts, 2023). In addition, President Joseph Biden proposed universal, publicly funded preschool for 3- and 4-year olds as part of the initial framework for his signature "Build Back Better Plan," (Popli and Vesoulis, 2021). Amid this nationwide expansion, we expect to see opportunities to learn more via randomization. This paper will help guide this effort by shedding light on what we do and don't know.

In practice, we elected to include any study where a substantial component of the intervention itself occurred during the typical preschool-age ranges in a classroom or group setting. This meant that birth-to-grammar school interventions like the classic Carolina Abecedarian Project and the more modern Educare were included as a focus of this review (Campbell and Ramey, 1995; Yazejian et al., 2017). However, work specific to

younger ages (such as the Infant Health Development Study) or that primarily occurred outside a classroom (such as Even Start) were not included (McCarton et al., 1997; Ricciuti et al., 2004).

1.3 Literature discovery protocol

We followed a modified version of the protocol outlined in Jackson and Mackevicius (2023) to find the papers that form the core focus of this review. We began by forming a set of "seed papers" that met our inclusion criteria by examining those included on three prominent research clearinghouse websites[3] as well as in the bibliographies of Gray-Lobe et al. (2023) and two recent early childhood education review papers (Cascio, 2021; Duncan et al., 2022). From there, we used Google Scholar to search for forward citations (i.e., newer papers that cite a paper in our seed corpus) and backward citations (i.e., older papers cited by a paper in our seed corpus). We then repeated this process of searching forward and backward citations until we found all the papers that met our inclusion criteria. In total, we arrived at 44 distinct papers describing results from 21 distinct experiments.

2 Earliest lottery evidence: randomized controlled trials from demonstration programs

Much of the evidence pertaining to the efficacy of preschool in the United States, and in particular its long-run effects, comes from a series of small-sample-size randomized controlled trials conducted in the 1960s and 1970s. In this section, we review these interventions and present novel pooled estimates of their impact across a range of outcomes over the life cycle. We also discuss some of the potential mechanisms that have been offered in the literature for the stark, and sometimes puzzling, findings from this body of work. We conclude by noting the limitations of the demonstration studies.

[3]These included the Center for Research on Children in the United States (www.crocus.georgetown.edu), the National Institute for Early Education Research (www.nieer.org), and the What Works Clearinghouse (www.ies.ed.gov/ncee/wwc/FWW).

2.1 Description of experiments

In this section, we provide details on each of the demonstration programs considered in this review. The programs are summarised in table 1, and summary statistics (where available) are given for the relevant populations in table 2.

These studies randomized a total of more than 350 children into preschool treatment and control conditions; had nearly universal take-up of treatment (likely a feature of the low-income populations being targeted and the lack of alternative preschool opportunities at the time); and showed remarkably low attrition in most follow-up surveys (often ≤ 10 percent), especially in the short-run (Anderson, 2008; Campbell et al., 2008; Conti et al., 2016; Heckman et al., 2010a).

Highscope / Perry Preschool (1962-1967).[4] Starting in 1962, the Perry Preschool Project (PPP) enrolled five cohorts of Black children in Ypsilanti, Michigan. To be eligible for the program, children had to exhibit a low IQ at baseline and score highly on a scale of cultural deprivation that included measures of parental schooling, the father's occupational status, and the quality of the family's housing (Conti et al., 2016). This disadvantage is reflected in the low paternal and maternal employment rates of the research subjects (see table 2). With the exception of the first cohort, Perry children enrolled in the program for two years beginning at age 3 (Schweinhart, 2004).

The intervention itself involved instruction in classes of five or six for five days a week (October-May) in 2.5 hour sessions for a total of 12.5 hours per week Schweinhart et al. (2005). The content of the classes was based on the work of Jean Piaget and emphasized language, socialization, numbers, space, and time. The intervention also included one 90-minute home visit per week (Anderson, 2008). Children randomized into the control group likely experienced a mix of neighborhood and at-home care (Conti et al., 2016).

The Early Training Project (1962-1964). Starting in 1962, the Early Training Project (ETP) enrolled two cohorts of 3- and 4-year-old Black children in Murfreesboro, Tennessee. The researchers focused on recruiting children whose parents had low levels of education, worked in low-income occupations, and lived in low-quality housing (Anderson, 2008; Klaus and Gray, 1968; Stevens Jr, 1982).

The intervention itself involved attending class in small groups of four-to-five chil-

[4]The year range for this description (and all subsequent intervention descriptions) denotes the time period of randomization.

dren, five days each week for 10 weeks during the year. The class emphasized positive reinforcement, motivation, and persistence (Anderson, 2008). Treated children also received one 90-minute home visit per week. The kinds of child care for children randomized into the control group were not clear from publicly available sources nor from the discussion in modern re-analysis of the data (Anderson, 2008). However, as with Perry, this time period pre-dates Head Start and, given that the sample of families was selected to be low on socio-economic indicators, it seems likely that the control group experienced a mix of home and neighborhood care.

The Carolina Abecedarian Project (1972-1977, 1978-1980). Starting in 1972, the Carolina Abecedarian Project (ABC) initially enrolled four cohorts of Black infants from Chapel Hill, North Carolina. Two additional cohorts of infants were enrolled in 1978-1980 in a follow-on study known as the Carolina Approach to Responsive Education (CARE), which was broadly similar to ABC including in eligibility criteria and overall scope of the intervention (Anderson, 2008; Campbell et al., 2008). For that reason, and because they are often analyzed together, we discuss them jointly here.

Infants enrolled in the program had to be free from any apparent adverse biological conditions and had to score highly on a composite risk index that included measures of parental education, parental IQ, family income, welfare receipt, paternal presence, stability of paternal employment, presence of maternal relatives, sibling school performance, social service receipt, and whether a family member sought counseling. Infants remained enrolled in the program until they reached schooling age (Anderson, 2008; Campbell et al., 2008; Conti et al., 2016).

The intervention involved attending preschool five days per week for eight hours per day in class sizes of typically six. The program emphasized cognitive, social, and language skills. The children enrolled in the Abecedarian project were also provided with on-site medical care that included immunizations, lab tests, health education, vision/hearing, and sick care. The control group was provided with free iron-fortified formula, free diapers, and social services (as needed). Because this program occurred after the launch of Head Start, it is likely that the control group children experienced a mix of in-home or neighborhood care along with various types of out-of-home care (Anderson, 2008; Campbell et al., 2008; Conti et al., 2016).

Table 1: Description of demonstration programs

Study	Intervention year	Location	Sample size	Inclusion criteria	Age	Intervention description	Control group	Childhood outcomes	Adult outcomes
Highscope / Perry Preschool	1962-1967	Ypsilanti, Michigan	123	• Child had low IQ • Cultural deprivation scale: parental schooling, father's occupational status, housing quality	3-4 years	• 5 days per week (until school age) • 2.5 hours per session • Class sizes of 5-6 • Weekly, 1.5 hour parent-teacher home visit	• In-home / neighborhood care	• IQ scores • In-school educational outcomes (e.g. grade retention, graduation)	• College attendance • Labor force participation • Criminal behavior • Marriage / family formation • Health • Second-gen outcomes
Early Training Project	1962-1964	Murfreesboro, Tennessee	65	• Low housing quality • Low parental education • Low income parental occupation	3-4 years	• 5 days per week (10 weeks a year over 2-3 years) • 4 hours per day • Class sizes of 4-5 • 1.5 hour home visit per week	• Unclear from primary sources, likely in-home / neighborhood care since this predates Headstart	• IQ scores • In-school educational outcomes (e.g. grade retention, graduation)	• College attendance • Labor force participation
Abecedarian Project	1972-1977 (ABC) 1978-1980 (CARE)	Chapel Hill, North Carolina	111 (ABC) 66 (CARE)	• No apparent biological conditions • High Risk Index: parental education, parental IQ, family income, welfare receipt, absent father, unstable paternal work record, presence of maternal relatives, sibling school performance, social services, family member sought counseling	4.4 months	• 5 days per week (until school age) • 8 hours per day. • Class size ~6 • On-site medical care (immunizations, lab tests, health ed, vision/hearing, sick care).	• Mix of in-home / neighborhood care + various types of out of home care. • Provided iron-fortified formula. • Free diapers • Supportive social services (as needed).	• IQ scores • In-school educational outcomes (e.g. grade retention, graduation)	• College attendance • Labor force participation • Criminal behavior • Marriage / family formation • Health

Notes: The information in this table is taken from descriptions contained in Anderson (2008), Conti et al. (2016), Heckman et al. (2010a,b), Klaus and Gray (1968), and Pungello et al. (2010)

Table 2: Demonstration program summary statistics

	Perry Preschool	Abecedarian	Early Training
Black	100%	98%	100%
Female	41.5%	53.2%	46.2%
Mother's age	25.6	19.8	
Mother's years of education	9.4	10.2	
Mother employed	20%	35%	
Father's age	32.8	23.21	
Father's years of education	8.6	10.9	
Father employed	14%	73%	
Father present	53%	29%	
Number of siblings	4.2	0.64	

Notes: Summary statistics for the demonstration programs are taken from Anderson (2008), Klaus and Gray (1968), and Conti et al. (2016). There are fewer summary statistics available in the published work about the Early Training Program, which is why they are blank in this table.

2.2 Findings

In this section, we summarize findings from the demonstration studies. Scholars have debated the reliability of the statistical inferences drawn from these programs due to the small sample sizes and potential for multiple testing. The latter possibility is particularly salient given the large number of outcomes considered and the fact that female and male subjects in the demonstration programs were often analyzed separately (see e.g., Anderson, 2008; Conti et al., 2016; Heckman et al., 2010b).

For that reason, we present de novo analysis of the publicly reported estimates that pools across genders and interventions.[5] Pooling across genders not only helps with power, but also limits the potential for inference to be incorrect due to multiple comparisons. Pooling across studies improves power and increases precision.

Wherever possible, we used the published estimates and standard errors from Anderson (2008) as the basis for the pooled averages. Anderson (2008) is the only published

[5]Precisely, we estimate the pooled average treatment effects using a maximum likelihood estimator: $\bar{\beta} = |\sum \frac{1}{SE_j^2}|^{-1} \sum \frac{\hat{\beta}_j}{SE_j^2}$ where j indexes a group-by-intervention cell (e.g. boys in Perry preschool), $\hat{\beta}_j$ is the corresponding cell-specific treatment effect, and SE_j^2 is the reported standard error.

paper that examines all three demonstration programs jointly, which has the advantage of allowing us to pool comparable estimates that have been otherwise treated similarly in the data processing and analysis steps that occurred prior to publication. In some cases, the nature of the available outcomes differs across studies (e.g., age of measurement or phrasing of the question). In all cases, we tried to find the closest match possible across studies. In cases where we could not closely approximate outcomes, we used whatever closely matching estimates were available (e.g., from one or two studies) to calculate a precision-weighted average. See appendix table A.1 for additional detail.

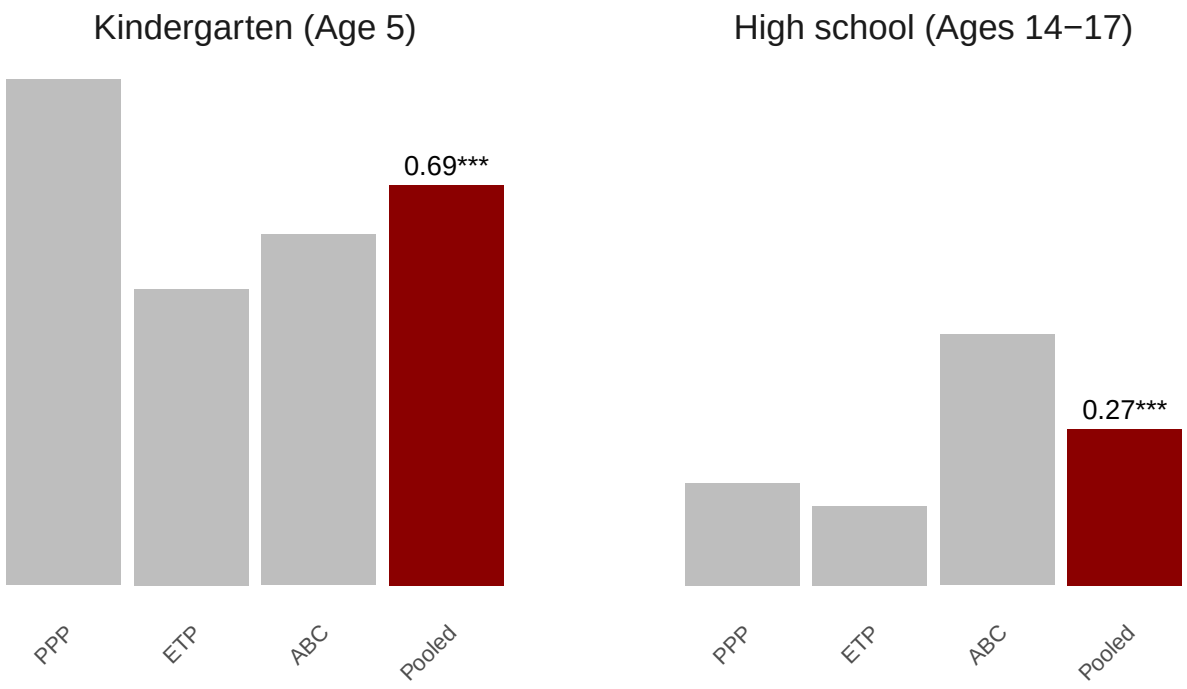
Demonstration program impact on IQ Scores. The impact on IQ scores merits close attention for several reasons. First, some version of an IQ score, psychometric scale, or state standardized test is available for virtually every study that warranted inclusion in this review. Thus, it forms a natural "common denominator" to benchmark the impact of different programs/interventions deployed in different times, settings, and study populations.

Second, the long-run outcomes related to overall social and economic well-being we sought to measure (e.g., labor force participation, educational attainment, and health) are typically impossible to observe until at least a decade or more has passed after an early childhood intervention has concluded. Some have argued that impacts on test scores can serve as a useful (if imperfect) short-run proxy for certain long-run outcomes of interest (e.g. Athey et al., 2019; Chetty et al., 2011, 2014). Others have argued that typical standardized test scores are inadequate either because they fail to capture important dimensions such as non-cognitive ability (Heckman et al., 2006) or because they can create a misleading impression of intervention longevity due to a statistical artifact of the way test scores are normalized (Cascio and Staiger, 2012). Thus, test scores merit special attention since a better understanding of their response to interventions that affect long-run outcomes is critical to understanding their usefulness as a short-run metric of intervention success.

Figure 1 displays the results. The left-hand panel shows impacts on intervention subjects at age 5, near school entry age. The right-hand panel displays treatment effects during the high school years. From left to right, the grey bars correspond to point estimates (pooled across genders) for Perry Preschool, the Early Training Program, and Abecedarian. The red bars display the corresponding pooled average.

The demonstration programs caused an enormous increase in short-run IQ scores. This was true for the individual programs, which ranged from 0.87σ (PPP) to 0.51σ (ETP) and all reached conventional levels of statistical significance. It was also true for the

Figure 1: Demonstration program effects on IQ scores



Notes: This figure displays consensus estimates aggregated from Anderson (2008) of the impact of the demonstration program interventions on IQ scores. The left-hand panel shows impacts on intervention subjects at age 5. The right-hand panel displays treatment effects during the high school years. From left to right, the grey bars correspond to point estimates for the Perry Preschool Program (PPP), the Early Training Program (ETP), and Abecedarian (ABC). The red bars display the corresponding pooled average. See appendix table A.1 for additional details on the construction of this figure, including the relevant point estimates and standard errors.

precision weighted average (0.61σ). To contextualize these effect sizes, a recent meta-analysis of charter school lottery studies found that the largest recorded effect size in existence for attending a charter school was 0.359σ (Chabrier et al., 2016).

However, evidence from the demonstration programs was mixed regarding the persistence of these impacts into the high school years. While the point estimates for all three programs remained large in absolute terms, they declined over time, ranging from 0.136σ (ETP) to 0.432σ (ABC). The individual program estimates were also imprecise and fail to reach conventional levels of statistical significance. Still, despite its decline, the pooled average *did* remain statistically significant, providing evidence of persistence that had (until now) gone unnoticed in the literature.

Demonstration program impact on later life well-being. A major advantage of the demonstration program studies is that enough time has elapsed between the intervention and the present to study the effects on later life outcomes.

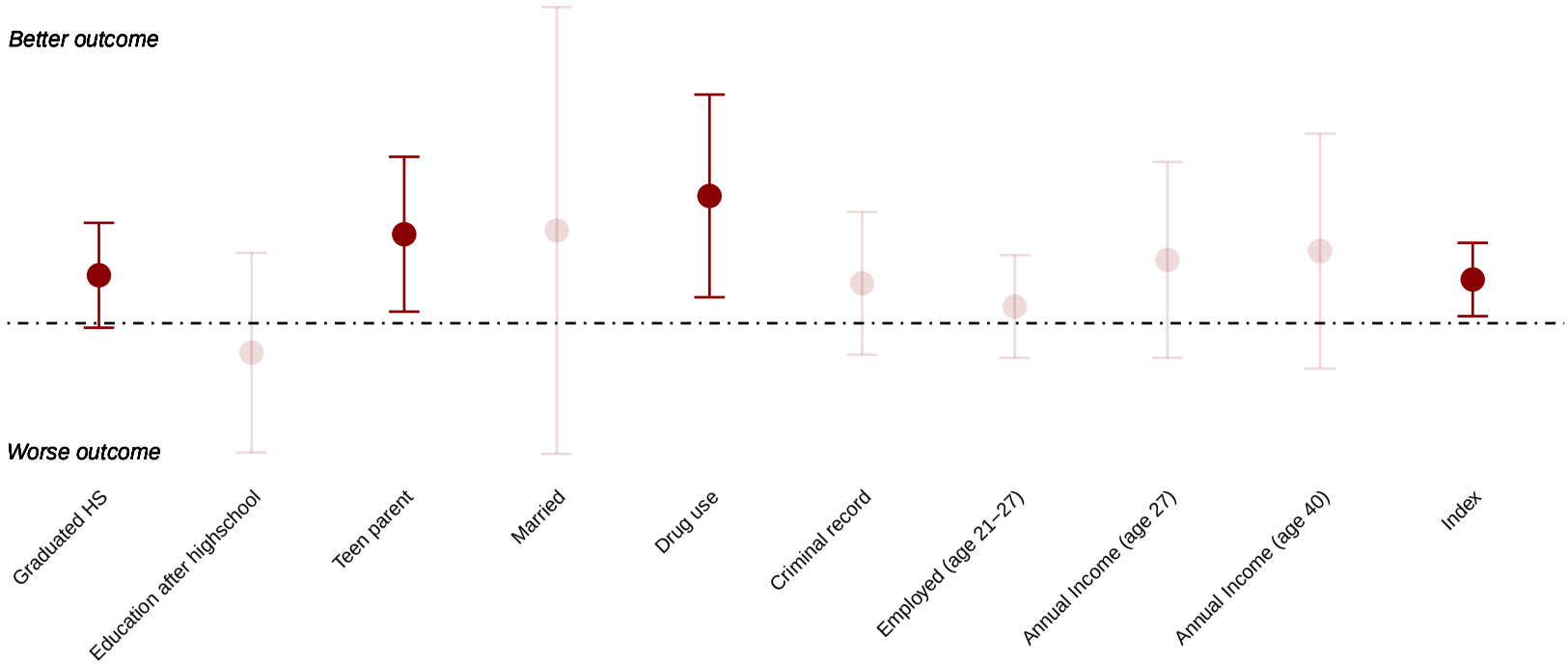
Figure 2 presents the results of the pooled analysis for a range of post-secondary outcomes. We selected these outcomes for meta-analysis because they were representative of the different *types* of well-being that have been explored in the demonstration program literature and because the measurement periods spanned a broad range of post-secondary ages. The milestones are arranged along the x-axis in the (rough) chronological order in which they tend to occur during the life cycle. To limit concerns related to multiple testing, we also present results for an index of adult outcomes constructed by Anderson (2008). With the exception of the adult index, the points represent pseudo-effect sizes[6] that scale the treatment effect as a share of the outcome mean.

The pooled estimates provide clear evidence that the demonstration programs were beneficial over the long run. Individually, we found statistically precise evidence that the programs improved the likelihood of high school graduation, reduced the likelihood of becoming a teen parent, and reduced the likelihood that the treated children engaged in the recreational use of illegal drugs. While the remaining outcomes were not estimated with enough precision to make strong statistical claims, we did find statistically precise evidence that the demonstration programs improved overall well-being as measured via the index of adult outcomes.

Mediation and non-cognitive skills. The fact that the test score effects fade over time

[6]Standard deviations necessary to calculate true effect sizes were not available for most of these outcomes in the published studies.

Figure 2: Demonstration program effects on later life outcomes



Notes: This figure presents consensus estimates aggregated from Anderson (2008) for post-secondary outcomes. Outcomes are arranged along the x-axis in the chronological order in which they tend to occur during the life-cycle. Points represent pseudo-effect sizes with treatment effects scaled as a share of the outcome mean. Whiskers represent 95 percent confidence intervals. Estimates that are statistically significant at the 10 percent level are bold and opaque; estimates that are not statistically significant are translucent. See appendix table A.2 for additional details on the construction of this figure, including the relevant point estimates and standard errors.

seems to suggest that the impact of the demonstration programs was transitory, perhaps as the control group "caught up" after entry into the K-12 system. Yet, the preceding analysis offers compelling evidence that the demonstration programs do affect long-run outcomes. What drives this pattern? Heckman et al. (2013) argue that in the case of Perry Preschool, the puzzle is explained by non-cognitive skills (as in Bowles and Gintis, 1976, 2001). They found that the program had persistent effects on externalizing behavior for men and women, which is defined as "aggressive, antisocial, and rule-breaking behaviors," (Heckman et al., 2013). They also found that the program had a persistent impact on cognition (men and women) and academic motivation (women only). This suggests that standardized test scores (and other cognitive measures) may be poor short-run proxies for the overall impact that preschool has on lifetime well-being, a pattern that will show up again when we discuss the modern studies in section 3.

Intergenerational effects. Most recently, researchers have explored the impact that the Perry Preschool Project had on the *children* of the original participants (García et al., 2023). They found that these children were more likely to grow up in stable two-parent homes and have parents who had above-average earnings, better health, and fewer run-ins with the criminal justice system. This appears to translate into downstream outcomes. Children of the original participants had higher levels of education, were more likely to be employed, had lower levels of criminal activity, and were healthier than the children of the control group.

However, childbirth itself does appear to be an outcome of the treatment, which makes the point estimates hard to interpret. While not significant individually in the Perry program, the effect sizes are large (e.g., for teen parenthood, the point estimates are 28 percent of the mean for females and 17 percent of the mean for males), and the more precise pooled average suggests that fertility is affected by the demonstration programs on average (see figure 2). The treatment effects on fertility therefore make these estimates challenging to interpret, because we cannot observe the counterfactual outcomes of children in the treatment group had they been born in an earlier cohort, nor can we observe the outcomes of children who were not born at all. This raises the possibility that the treatment effects are not generated by changes in the nature of intergenerational investments, but are actually the product of secular trends or simply biological benefits due to having older parents.

Limitations. The demonstration programs have undoubtedly been influential, serving

as the basis of both a large academic literature and as the foundation for calls to expand preschool in the United States. However, this literature has clear and important limitations.

First, the demonstration programs were of exceptionally high quality and intensity and, in many ways, look nothing like the current landscape of preschool in the United States. For example, Perry Preschool and the Early Training Project both incorporated home visitations and direct parental engagement into their curriculum. At the most extreme, the Abecedarian program included high-quality, on-site medical care for the children. In all three cases, the group size was extremely small (four-six children per classroom); current federal regulation/guidelines call for 15-17 children per classroom in Head Start Centers (*U.S. Code Title 45 - Public Welfare*, n.d.).

Second, in all three cases, the subjects enrolled in the study were extremely low-income and nearly 100% Black. This leaves open the possibility that the results may not extend to more affluent populations who might begin preschool at a different baseline or have access to other forms of childhood investment. It is also possible that the unique barriers and challenges faced by the Black community in the United States may affect the applicability of the findings.

Finally, we note that half a century has passed since the initial randomization occurred. The social, economic, and political landscape of the United States is fundamentally different today than it was at the time these experiments were conducted. More concretely, the prevalence of preschool has grown dramatically over this period, with enrollment rates for 4-year-olds climbing from nearly zero in 1950 to 67.7 percent today (US Department of Education, 2019). As a result of this growth, and in particular, the expansion of subsidized programs like Head Start, the "outside option" of even the lowest income families is likely to be dramatically different than it was 50 years ago (Cascio, 2021; Duncan and Magnuson, 2013).

As a result of all these limitations, the estimates from the demonstration programs are best thought of as a proof of concept yielding "upper bounds" on what is possible to achieve with intensive and high-quality childhood educational interventions, rather than as a realistic cost-benefit assessment applicable to the decisions under consideration today.

3 Modern lottery evidence on preschool attendance

We now summarize findings from modern studies that involve larger samples than the older ones and that use a research design involving a randomized offer to enter a preschool program. We begin by reviewing the details of the programs themselves and the nature of the experiments. We then discuss their short- and medium-term findings. We conclude by discussing nuances and limitations of this evidence.

3.1 Description of experiments

In this section, we describe the details of the experiments considered. The programs are summarized in table 1 and demographic characteristics of the relevant populations are provided in table 2.

Unlike the demonstration studies, most of the experiments we explore in this section involve large samples and hence do not suffer the issues related to power that affect the demonstration studies. Other than standardized tests / IQ scores, the outcomes collected in these studies exhibit much less overlap than those explored in the demonstration programs. For those reasons, we will not present pooled estimates and instead review treatment effects on various outcomes program-by-program.

The modern lottery evidence also differs from the demonstration programs in another important aspect: the degree of takeup. While the demonstration programs saw nearly universal acceptance of offers to attend an early childhood experience, many families in the modern studies declined randomly assigned offers to attend preschool. Some families found a way to secure a spot in the program despite being denied a randomly allocated offer.

Two statistical approaches are generally accepted to handle these complications related to selective takeup of treatment. The first would be to focus on "intent to treat" or ITT estimates, which capture the effect of the randomly assigned *offer* to enroll in a preschool, rather than the causal effect of the preschool itself. The second would be to focus on estimates that correct for non-compliance by scaling up the ITT by the estimated share of compliers in the data (e.g., instrumental variables, two-stage least squares, and the methods developed in Abdulkadiroğlu et al. (2017)). The estimates derived from the latter approach would be interpreted as "Local Average Treatment Effects," or LATEs, and they represent the causal effect of the treatment for the sub-population of study subjects that comply with lottery offers.

While both approaches have advantages and disadvantages, we prefer the LATE estimates since they are more directly comparable to the results from the demonstration studies. In practice, the studies in this section that met our inclusion criteria more commonly and consistently reported LATE estimates than the corresponding ITT. As a result, we focused on the LATE estimates exclusively.

We now provide short descriptions of these modern interventions: populations of study, intervention years, and the nature of the preschool program attended.

Boston universal preschool (1997-2003, 2007-2011). Boston Public Schools has operated a large preschool program since the 1990s. The classrooms are located in school facilities and early learning centers and, during the study period, were staffed by certified BPS teachers who held either a bachelor's or a master's degree. Fifty-six percent of teachers held master's degrees, and the average teacher had eight years of experience. During the study period, the average class size was 19 students. Per-student funding for the BPS program was above what is typically received by Head Start and nearly double the average state-funded program. The program was initially half day and transitioned to full day over the course of the cohorts studied in Gray-Lobe et al. (2023). The preschool curriculum used by the district also varied over this time frame, starting with Harcourt Trophies and transitioning to Opening World of Learning and Building Blocks (Gray-Lobe et al., 2023).

Any 4-year-old student who resides in Boston is eligible to enroll in publicly funded preschool. In practice, demand exceeds capacity. As a result, the district uses a random number generator to allocate seats among families with similar preferences and priorities within the centralized school assignment mechanism. This provides the random variation in treatment assignment that Gray-Lobe et al. (2023) and Weiland et al. (2020) use to generate causal estimates of the impact of preschool and is conceptually similar to the research designs of the other studies explored in this section that randomly allocate offers among applicants to oversubscribed programs. The primary methodological difference between Weiland et al. (2020) and Gray-Lobe et al. (2023) is that the former uses a "first-choice" design, whereas the latter uses estimation techniques designed to leverage fully the random variation embedded in the system; for that reason, we focused on treatment effects reported in Gray-Lobe et al. (2023) whenever they were available. Substantively, Weiland et al. (2020) focuses on outcomes through grade three for more recent cohorts (2007-2011), whereas Gray-Lobe et al. (2023) focuses on longer-run outcomes for earlier cohorts. In practice, the students who were not offered a spot in the program were likely

Table 3: Description of modern extensive margin lotteries

Study	Intervention year	Location	Sample size	Inclusion criteria	Age	Outcomes
Boston universal pre-k	1997-2003 2007-2011	Boston, Massachusetts	4215	Applied to attend pre-k in Boston and was ``at-risk'' of lottery tie-breaking due to preferences and priorities in school assignment mechanism.	4	• Standardized test scores • School outcomes (grades, retention, special education services, attendance, discipline) • Graduation & college enrollment • Criminal justice involvement
Head Start Impact study	2002-2003	National	4385	Applied to, and was qualified for a seat in, a local Headstart program at one of 340 sites across the US.	3-4	• IQ Scores • Parental involvement • Parental observations of child behavior
Tennessee statewide pre-k	2009-2011	Tennessee (statewide)	2990	Students had to be eligible for free or reduced price lunch, four years old, applicants to an oversubscribed Tennessee Pre-k program, and not apply for special education services prior to enrollment	4	• Standardized test scores • Special education services • School discipline
Educare	2010	Chicago, Illinois; Milwaukee, Wisconsin; Omaha, Nebraska; and Tulsa, Oklahoma.	206	Families applied for a spot in one of five Educare child care centers and met the usual head start poverty criteria. In addition, to be eligible, the child had to be less than 19 months in age and not an enrolled child's sibling, a staff child, or a foster child. From this pool, the researchers attempted to recruit the "highest risk" families.	<19 months	• IQ Scores • Parent-child interactions • Parental observations of child behavior

Notes: The information in this table is taken from accounts in Durkin et al. (2022), Feller et al. (2016), Gray-Lobe et al. (2023), Puma et al. (2010), Walters (2015), Weiland et al. (2020), and Yazejian et al. (2017)

drawn (in roughly equal proportions) from a mix of Head Start centers, private providers, and home or neighborhood care (Gray-Lobe et al., 2023).

Head Start (2002-2003). Head Start is the largest early childhood education program in the nation. Since its inception in 1965, it has served more than 38 million children, with more than 750,000 enrolled in 2020 alone at a total cost of over $10 billion. Head Start is a year-round program designed to promote school readiness for low-income families (US Department of Health and Human Services, n.d.). Class sizes for 4- and 5-year-old children are capped at 20. The cap for classes serving 3-year-olds is 17 *U.S. Code Title 45 - Public Welfare* (n.d.). Head Start programs also screen children for developmental delays and engage families with home visits, parenting classes, and other social services. The length of the school day varies from site to site.

The Head Start Impact Study was a randomized evaluation of the program that emerged from the 1998 Head Start reauthorization act (Puma et al., 2010). The research design involved randomizing enrollment offers among applicants at oversubscribed Head Start centers. Ultimately, more than 4,000 children from 353 distinct Head Start centers within 84 regional programs were involved in the randomization and subsequent data collection (Feller et al., 2016; Walters, 2015). Depending on the age and enrollment cohort, from 40 to 87 percent of the children not offered a spot in a local Head Start center either enrolled in an alternative Head Start center or obtained some other form of center-based care (Kline and Walters, 2016).

Tennessee state-wide preschool (2009-2011). Tennessee has operated a state-funded preschool program for low-income families since 2005. It serves more than 18,000 4-year-old children. The instructional day lasts a minimum of 5.5 hours and runs for five days each week. Classes are capped at 20 students and must be staffed by a state-licensed teacher. Each room must have an educational assistant. Sites select a curriculum from a pre-approved list (Durkin et al., 2022).

As in the Head Start Impact Study, the research design involved randomizing offers to enroll at oversubscribed preschool sites. In practice, to be included in the sample, the children had to be eligible for free- or reduced-price lunch, be 4 years old by the end of September during their preschool year, and not have previously applied for special education services (outside the classroom). The sample ultimately included nearly 3,000 students across 79 sites and over two cohorts (2009-2010 and 2010-2011).

Table 4: Modern lottery summary statistics

	Boston	HSIS	TVP	Educare
Black	41%	30%	20%	57%
Hispanic	34%	37%	13%	32%
Female	49%	49%	51%	55%
Spanish language	19%	27%	13%	36%
Age (years)	4.6	3.5	4.4	0.8

Notes: The summary statistics in this table are taken from Durkin et al. (2022), Feller et al. (2016), Gray-Lobe et al. (2023), Kline and Walters (2016), and Yazejian et al. (2017).

Educare (2010). Educare is an enhanced early Head Start program that serves children from low-income families from birth to age 5. Educare classrooms have a teacher-to-student ratio of three to eight. All teachers must have a bachelor's degree. The school day runs 8-10 hours per day, five days per week, with all children required to attend for at least six hours a day. The educational model emphasizes data-driven approaches, teacher professional development, and strong partnerships between families and schools. The program operates across 21 sites in 18 cities (Yazejian et al., 2017).

The study recruited participants across five Educare sites located in Chicago, Milwaukee, Omaha, and Tulsa. The research design involved making random offers of attendance among families who had applied for oversubscribed slots. To be eligible for the study, families had to meet the usual Head Start poverty criteria, and the child had to be younger than 19-months-old, not the sibling of an enrolled child, and not a foster child. From this pool, the researchers write that they attempted to recruit the "highest risk" families; however, the criteria used to determine high risk is not clearly explained in the published work (Yazejian et al., 2017).

3.2 Findings: short-term outcomes

Test scores. As with the demonstration programs, we observe large impacts on short-run test scores that appear to fade out over time. Figure 3 plots test score effect sizes (in standard deviation units) from the four programs considered in this section and at the (approximate) age when the child was tested. The color of the points denotes the intervention. The diameter of the points is proportional to precision, so that larger dots

represent more exact estimates. Estimates that are statistically significant at the 95 percent level are bold and opaque; estimates that are not statistically significant are translucent.

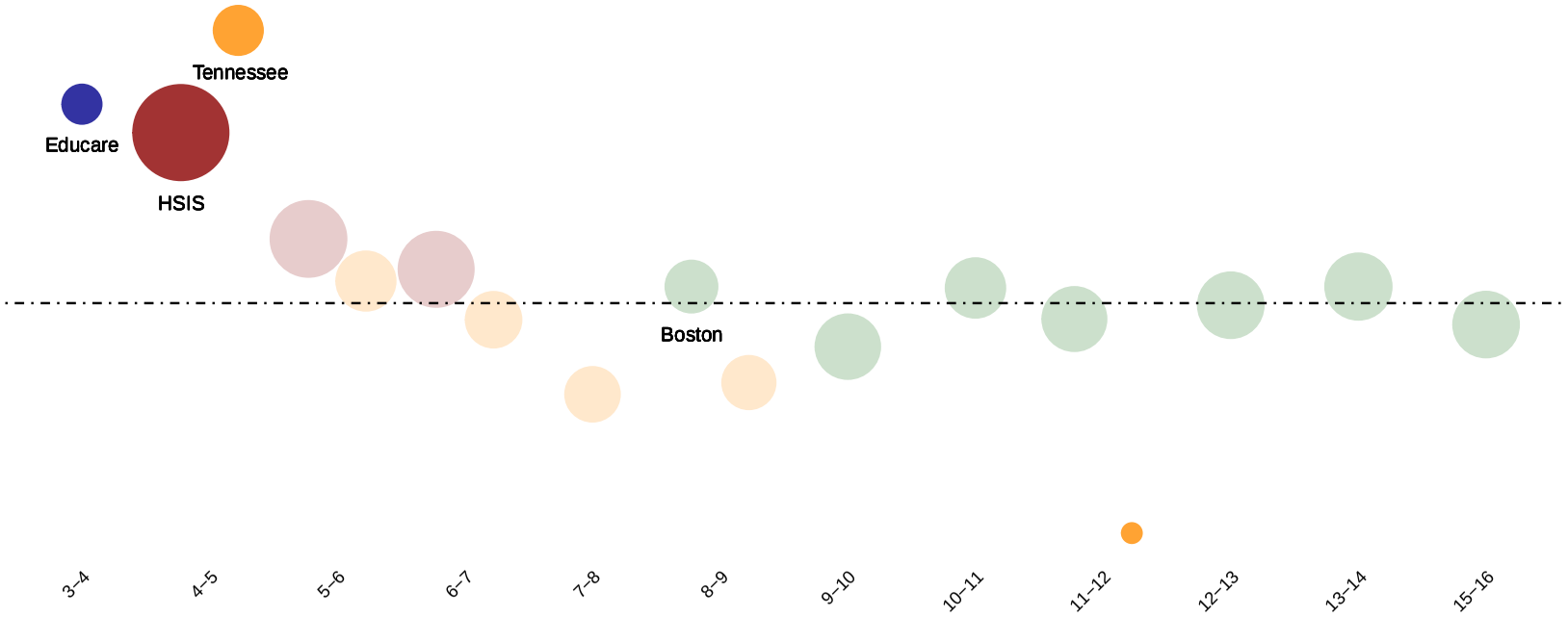
The estimates for the 3-5 age groups represented in figure 3 range from $0.25 - 0.395\sigma$. These effect sizes are generally smaller than the point estimates for the demonstration programs. That said, it is important to note that in absolute terms they are still large in magnitude.

Why are the short-run impacts smaller for the modern programs? Two papers found that the (relatively) small short-run effect sizes for Head Start were, in part, a product of the counterfactual preschool experiences of the control group children (Feller et al., 2016; Kline and Walters, 2016). Unlike the demonstration studies, where the control group children largely experienced a mix of at-home and neighborhood care, a large portion of the children from the control families in the Head Start Impact Study attended other, center-based care, including competing Head Start programs. Both Feller et al. (2016) and Kline and Walters (2016) found that the treatment effect was much larger for families that complied with lottery offers by moving their children out of home-based care, with effect sizes potentially as high as 0.49σ and well within the range of the demonstration program estimates after accounting for statistical uncertainty. These findings highlight the importance of considering the counterfactual for program targeting when deciding whether (and by how much) to scale the modern early-childhood education system.

Beginning at ages 5-6, we also see substantial evidence of fade-out similar to the patterns found in the demonstration program studies. This pattern is true whether we look within intervention (as is possible with the Head Start Impact Study and Tennessee statewide preschool) or take the totality of the evidence across interventions (such as comparing Boston and Educare, which only observed test scores in the later / earlier years respectively). In fact, the most recent follow-up to the Tennessee intervention found statistically significant *negative* point estimates, suggesting that access to preschool might have actually hurt these children on average. However, these findings from Tennessee should be treated with a great deal of caution. As figure 3 makes clear, this outcome was not only an outlier within the distribution of evidence, but it was also the least precisely estimated effect out of all the studies considered. This raises the very real possibility that the cause is sampling variation rather than a real treatment effect. On balance, the evidence in figure 3 suggests clear short-run improvements followed by rapid fade-out such that there is no statistically significant evidence of a medium- or long-run effect.

Academic performance, school discipline, and anti-social behavior. Evidence is mixed

Figure 3: Modern preschool: impact on test scores



Notes: This figure plots test score effect sizes (in standard deviation units) from Boston, Head Start, Tennessee, and Educare at the (approximate) age when the child was tested. The color of the points denotes the intervention. The diameter of the points is proportional to precision, so that larger dots represent more exact estimates. Estimates that are statistically significant at the 10 percent level are bold and opaque; estimates that are not statistically significant are translucent. See appendix table A.3 for additional details on the construction of this figure, including the relevant point estimates and standard errors.

from the modern period concerning the impact of preschool on outcomes measured during school-aged years. Unlike the research elsewhere in the modern period, the Boston and Tennessee studies could match participants to administrative data from the K-12 education system. This allowed the authors to construct estimates of the program's impact on non-test score measures of academic outcomes and behavior, including school discipline and, in the case of Boston, measures of juvenile incarceration. These outcomes are significant because the stated purpose of public funding for Head Start and many other publicly funded early-childhood education programs is to ensure children arrive at the K-12 system with the skills they need to be successful.[7]

Figure 4 presents results for various measures of non-test score-based academic performance. The x-axis displays outcomes grouped by grade level. The points represent pseudo-effect sizes[8] that scale the treatment effect by the control group outcome mean and the whiskers represent 95 percent confidence intervals. Colors denote the intervention.

Across a wide range of measures, we see little evidence from either intervention that these preschool experiences caused meaningful changes in overall academic performance and preparedness. For example, treatment effects on teacher-child observation ratings from the Tennessee study meant to capture the academic preparedness of students entering kindergarten were near zero (0.06 and -0.02 respectively) and estimated with enough precision to rule out meaningful changes. Similarly, in Boston, the authors found no effect on absenteeism in middle and high school. The one outcome with statistically significant and large effects was from the Tennessee study, which found that treated students were more likely to be diverse learners as defined by having an Individual Education Plan (IEP) in middle school. However, a similar outcome measured during elementary school from the Boston study examining more recent cohorts found no such effect (Weiland et al., 2020). Thus, the balance of the evidence suggests that these interventions had little impact on non-test score-based measures of academic performance.

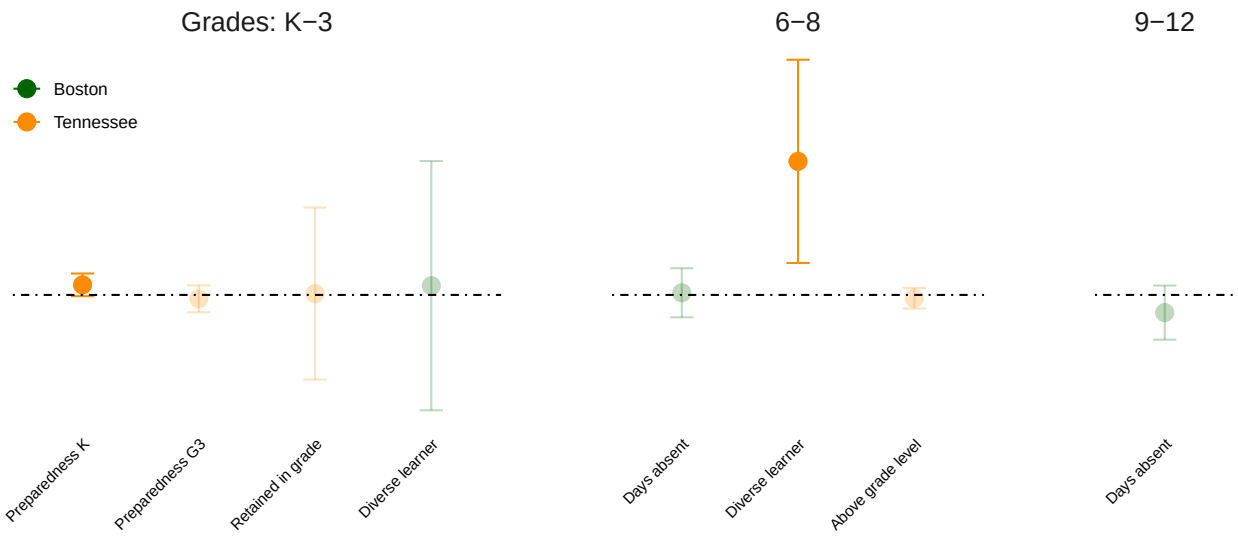
Figure 5 presents results for school disciplinary measures and juvenile incarceration. The x-axis displays outcomes grouped by grade level. The points represent pseudo-effect sizes[9] that scale the treatment effect by the control group mean and the whiskers represent

[7]For example, federal government websites (e.g. www.benefits.gov/benefit/1937) describe the program as follows: "Head Start is a Federal program that promotes the school readiness of children from birth to age five from low-income families by enhancing their cognitive, social, and emotional development."

[8]Standard deviations necessary to calculate true effect sizes were not available for most of these outcomes in the published studies.

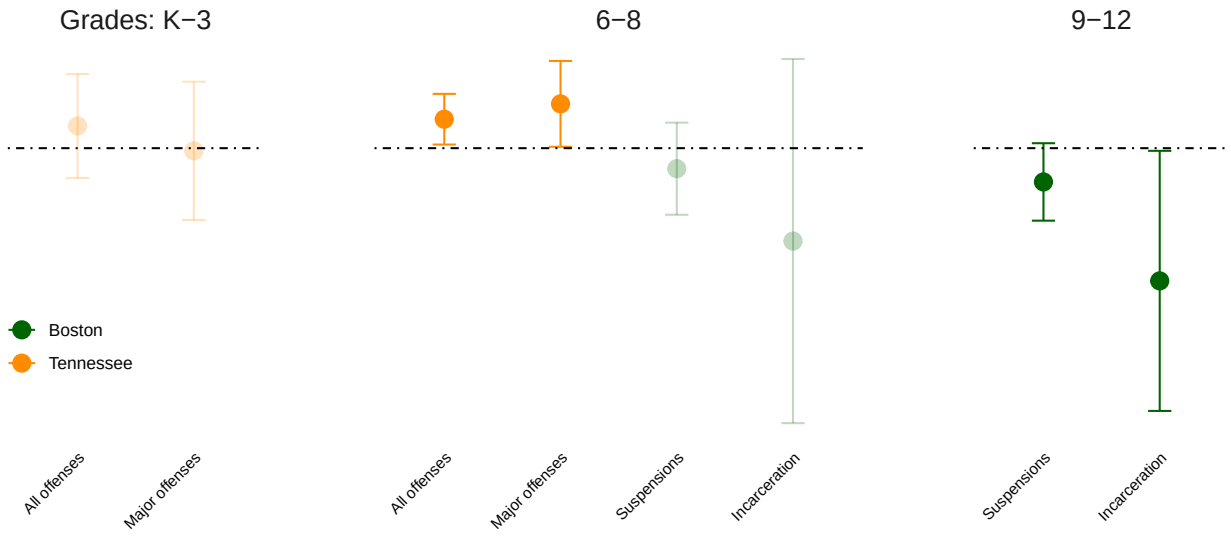
[9]Standard deviations necessary to calculate true effect sizes were not available for most of these outcomes

Figure 4: Modern preschool: impact on academic performance



Notes: This figure presents results for various measures of non-test score-based academic performance. The x-axis displays outcomes grouped by grade level. The points represent pseudo-effect sizes that scale the treatment effect by the control group outcome mean, and the whiskers represent 95 percent confidence intervals. Estimates that are statistically significant at the 10 percent level are bold and opaque; estimates that are not statistically significant are translucent. See appendix table A.4 for additional details on the construction of this figure, including the relevant point estimates and standard errors.

Figure 5: Modern preschool: impact on school discipline and anti-social behavior



Notes: This figure presents results for school disciplinary measures and juvenile incarceration. The x-axis displays outcomes grouped by grade level. The points represent pseudo-effect sizes[10] that scale the treatment effect by the control group mean and the whiskers represent 95 percent confidence intervals. Estimates that are statistically significant at the 10 percent level are bold and opaque; estimates that are not statistically significant are translucent. Colors denote the intervention. See appendix table A.5 for additional details on the construction of this figure, including the relevant point estimates and standard errors.

95 percent confidence intervals. Colors denote the intervention.

The disciplinary data yield stark and contradictory findings. On the one hand, Gray-Lobe et al. (2023) found small reductions in the likelihood a student was suspended. They also found large reductions in juvenile incarceration. However, among these outcomes, only juvenile incarceration during high school meets conventional levels of statistical significance. On the other hand, Durkin et al. (2022) found small but statistically precise evidence of an increase in disciplinary violations, and in particular major violations, happening in middle school. What drives these seemingly contradictory findings?

One possibility is that these results represent some form of real heterogeneity in the quality of preschool, the populations of study, or the counterfactual across the two environments. However, this explanation is puzzling in light of the short-run benefits the Tennessee study found for test scores (which echo the findings of nearly every other random-

in the published studies.

ized evaluation of the causal effect of US preschool, and which nearly universally show medium-term benefits on similar outcomes). Another possibility is that these findings simply represent statistical noise. However, unlike the anomalous test score outcomes discussed in figure 3, the comparable estimates from Boston offer levels of precision similar to those from Tennessee.

A third possibility relates to an important caveat that applies to all of the findings from the in-school outcomes explored in the Tennessee and Boston studies.[11] All of these papers found that the offer to enroll in preschool appeared to cause a small reduction in study attrition (Gray-Lobe et al., 2023; Durkin et al., 2022; Weiland et al., 2020). In other words, treated children in these studies were more likely to be found in the school administrative record system than the control group children. This differential attrition would skew the estimated treatment effects if the outcomes of children who were missing tended to be better (or worse) on average than the children matched to the state data. Thus, it could be the case that one (or both) of these sets of results reflect some bias resulting from the differential attrition.

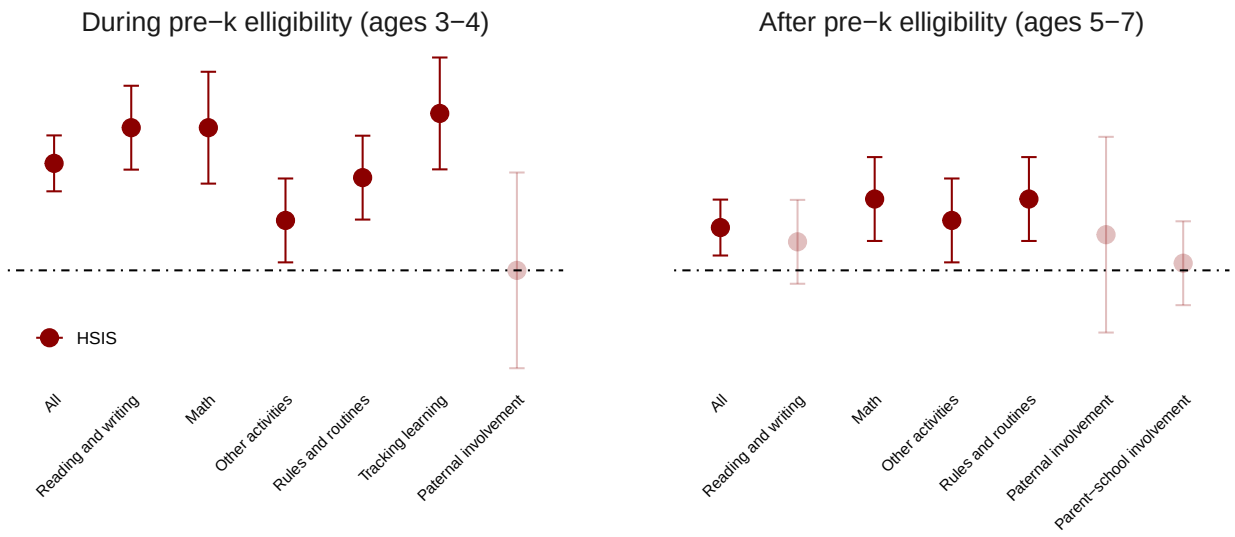
However, in the case of Gray-Lobe et al. (2023), the authors were also able to match their sample to administrative data with national coverage on college attendance (discussed in more detail in section 3.3). They found that the attrition problem was much smaller in this sample. As discussed in more detail in section 3.3, the authors also found benefits to college attendance that were broadly consistent with the effects they documented on disciplinary violations and anti-social behavior using the state administrative data. So while not definitive on the question of differential attrition, the Boston findings are consistent with the overall conclusion that the program benefited students during their school-age years.

Parental engagement. Evidence from the Head Start impact study suggests that high-quality preschool can cause parents to increase their level of engagement with their children. The Head Start Impact Study collected rich data on parental involvement both during and after the conclusion of the preschool experience.

Gelber and Isen (2013) use this survey data to construct indices meant to represent different domains of parent-child interaction. Figure 6 reproduces their published point estimates for the subset of domains that are most closely aligned with the cognitive and non-cognitive channels that have received attention in the literature from the demonstra-

[11] This caveat also applies to the test score-based finding from these two studies in the preceding section

Figure 6: Modern preschool: impact on parental engagement

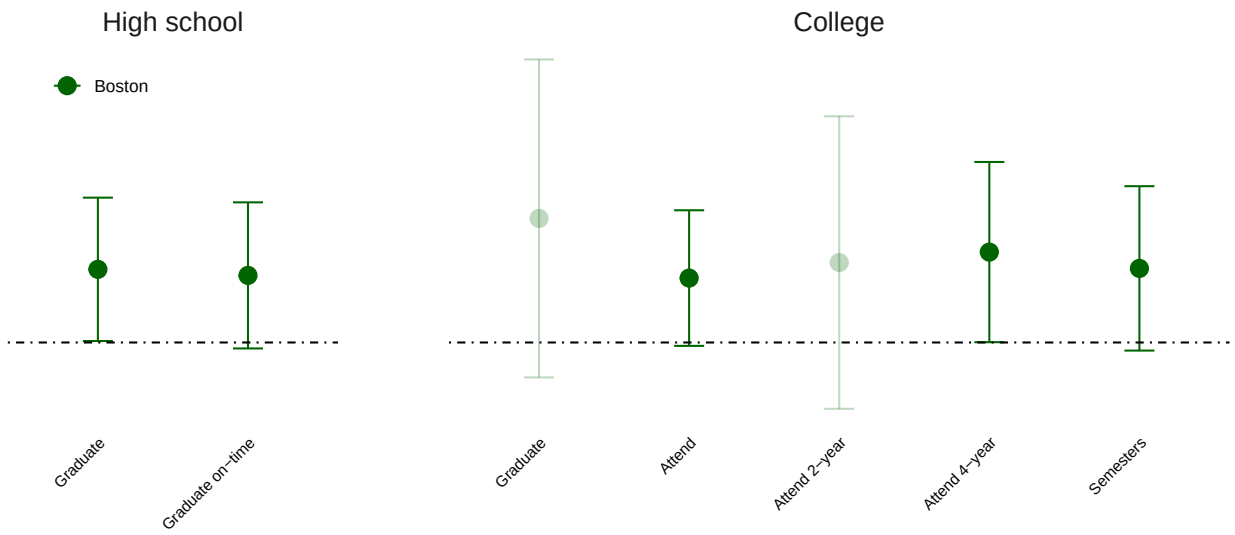


Notes: This figure reproduces published point estimates from Gelber and Isen (2013). Treatment effects are measured in standard deviation units, and the whiskers represent 95 percent confidence intervals. Estimates that are statistically significant at the 10 percent level are bold and opaque; estimates that are not statistically significant are translucent. See appendix table A.6 for additional details on the construction of this figure, including the relevant point estimates and standard errors.

tion studies. The points themselves are measured in standard deviation units, and the whiskers represent 95 percent confidence intervals.

The authors found that Head Start increased parental engagement along a variety of domains. During the preschool-eligible age years, parents in the treatment group spent more time on activities related to cognitive skills such as reading and writing and tracking their child's learning. Parents also were more engaged in areas that are plausibly related to non-cognitive development such as practicing rules and routines. These effects also appear to be long-lasting. While the point estimates were considerably smaller over time, the authors found that the increase in parental engagement lasted even after the preschool experience had ended. Thus, these impacts appear to represent real, lasting changes and not a "mechanical effect" of the high-quality Head Start curriculum prompting behavior changes that disappear once the child ages out of preschool and into a standard K-12 curriculum.

Figure 7: Modern preschool: impact on educational attainment



Notes: This figure displays effect sizes related to medium-term outcomes. The x-axis displays outcomes grouped by whether they pertain to high school graduation or college-going. The points represent pseudo-effect sizes with treatment effects scaled by the control group mean. The whiskers represent 95 percent confidence intervals. Estimates that are statistically significant at the 10 percent level are bold and opaque; estimates that are not statistically significant are translucent. See appendix table A.7 for additional details on the construction of this figure, including the relevant point estimates and standard errors.

3.3 Findings: outcomes in the medium term

Evidence from the modern period suggests that preschool has a beneficial impact on post-secondary outcomes like high school graduation and college attendance. Figure 7 displays results from Gray-Lobe et al. (2023), which is the only study from the modern period that explores the impact of preschool on post-secondary outcomes. The x-axis displays outcomes grouped by whether they pertain to high school graduation or college-going. The points represent pseudo-effect sizes[12] with treatment effects scaled by the control group mean. The whiskers represent 95 percent confidence intervals.

Across all outcomes explored in table 7, the effect sizes are moderate (ranging from 8 to 16 percent of the control group mean). For high school graduation and four-year college attendance, the treatment effects are also statistically significant at conventional levels. These point estimates suggest that attending a Boston preschool site increased a

[12]Standard deviations necessary to calculate true effect sizes were not available for most of these outcomes in the published studies.

child's probability of graduating high school by 6 percent and their probability of four-year college attendance by 5.9 percent. Taken together, the results provide clear evidence of beneficial medium-term effects.

3.4 Discussion

Comparison to the results from demonstration programs. The evidence from modern programs that vary access to preschool experiences echoes many of the big-picture patterns found in the demonstration studies. For example, both the demonstration program literature and more modern work found clear evidence of short-term cognitive effects that faded upon entry into the K-12 system. The one study where adult outcomes were available offered clear evidence of long-run effects.

However, the balance of the evidence differs in important ways. For example, the modern studies provide a mix of null and contradictory results on non-test score-based juvenile outcomes (like disciplinary actions and school preparedness) that could speak to the non-cognitive channels believed to be responsible for the long-run effects among demonstration study participants. Effect sizes are generally smaller among the modern programs. As discussed in detail in section 3.2, this is in part due to the changing nature of the counterfactual, with the demonstration study participants drawn from neighborhood / home care and the modern studies pulling a substantial share out of some alternative, center-based early childhood education experience. However, as noted in section 2.2, the demonstration studies were of exceptional quality for their time and by today's standards, and they targeted exceptionally disadvantaged populations. These factors raise the possibility that at least part of the attenuation relative to early research is due to the fact that the modern work generally explores the effect of preschool attendance on real-world programs serving more representative samples at scale.

Variation in site-level impact and the importance of intervention quality. Variation in treatment effects across Head Start sites is the subject of Walters (2015). Unlike prior work on Head Start that explored the treatment effect from an average Head Start experience across the 340 sites included in the randomization, Walters (2015) recognizes that the substantial variation in the characteristics of these sites may impact their relative effectiveness. The author treats each site as a separate experiment, allowing him to ask: "What are the site-level characteristics that predict a high-quality Head Start experience?"

Walters (2015) found substantial variation in quality, with the standard deviation of

site-level effects reaching 0.18σ. To contextualize this finding, it implies that a Head Start site that is two standard deviations above average improves the test scores of the students who attend by 0.48σ relative to an average effect of 0.11σ (using the author's preferred estimates). The author found that Head Start sites that offer full-day services and home visitation were more effective; whereas the High/Scope curriculum, better-educated teachers, and class size were not were not significantly different than average. However, these findings come with an important caveat. Just knowing the general characteristics of an effective preschool is not enough to identify precisely what makes it successful; other, less observable, factors may be at play.

Thus, these findings beg the question: "What practices do cause site-level improvements in preschool quality?" As Cascio (2021) and Duncan and Magnuson (2013) point out, while preschool is not yet universally guaranteed in the United States, enrollment has dramatically increased over the past 50 years. Policymakers today are not just focused on increasing access but also on identifying and scaling up the most effective educational practices.

4 Lottery evidence on drivers of quality

We now summarize findings from modern studies that explore the drivers of quality by randomly varying characteristics, policies, and practices across preschool providers. Before exploring these results, we note that much of this evidence comes from papers published in psychology journals; consequently, important methodological differences should be considered when comparing this body of work to the studies reviewed in the prior sections. First, the default choice for estimation in this literature is hierarchical or multi-level modeling. Such an approach can improve precision when the additional structure required by these models is met in the data (Gelman and Hill, 2006; Hansen, 2022). However, beyond the multi-level structure, these papers do not typically account for statistical dependence across observations, even in cases where treatment assignment and/or the sampling structure clearly operate in clusters. This leaves open the possibility that the inferences in this literature are not robust to heteroscedasticity or forms of clustered error dependence that are common in applied work. It is also common in this literature to report statistical significance without including the corresponding p-value or standard error, which makes more in-depth forms of meta-analysis impossible. Finally, we note that this literature often presents estimates without formally testing for sample

balance on predetermined characteristics, a standard practice in randomized evaluations to confirm the validity of the research design.

We present this evidence even with these caveats because so few studies have explored random variation in preschool characteristics. However, as a result of the limitations, we focus attention on effect sizes and broad patterns across studies rather than on specific inferences drawn from any one experiment.

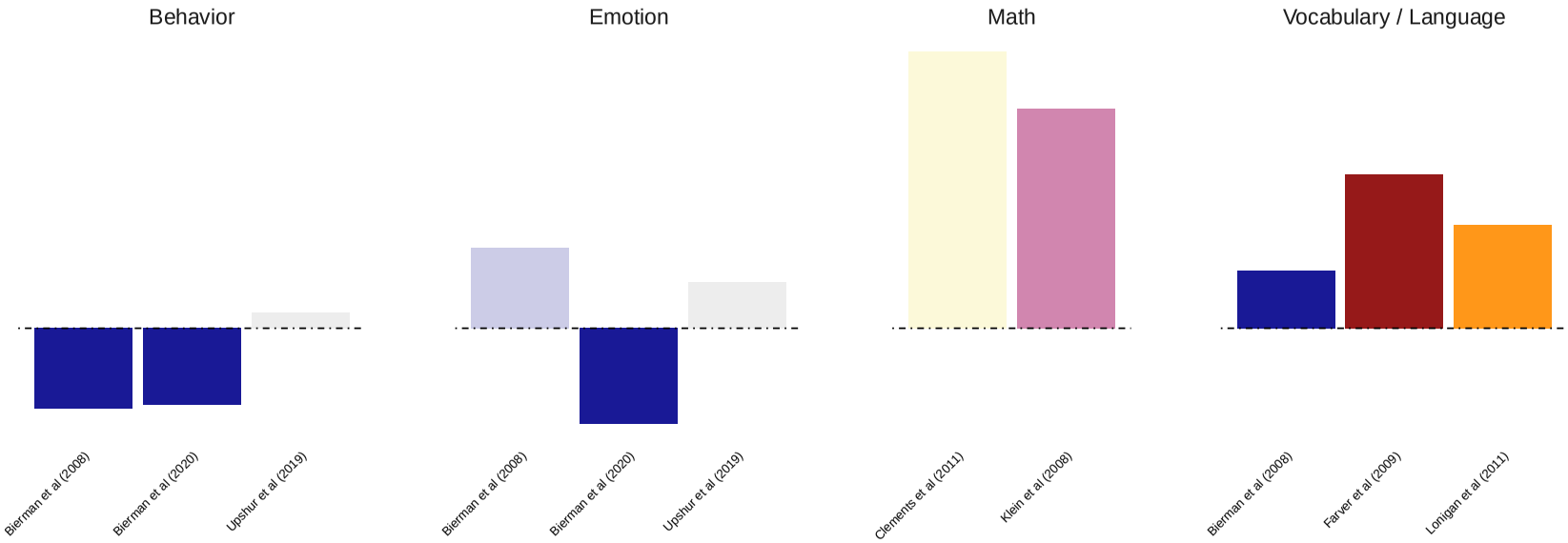
4.1 Curriculum

Evidence suggest that specific curriculum can improve emergent academic skills such as math and vocabulary. However, results for behavioral and social-emotional skills are more mixed. In total, our search uncovered six studies that randomly varied the curriculum used in the classroom. Figure 8 plots effect sizes from these studies organized by the broad developmental area targeted by the intervention.

The two studies we uncovered that explore mathematics-focused curricula found promising results supported by well-done RCTs with reasonably large sample sizes. Clements et al. (2011) study the Building Blocks mathematics curriculum by randomizing the use of the curriculum across 106 classrooms. The curriculum involves guiding children "to extend and mathematize their everyday activities, from block building to art to songs to puzzles, through sequenced, explicit activities" and incorporates computer-based elements (Clements et al., 2011). The researchers found large effects (0.72σ) on an early childhood math assessment. Klein et al. (2008) studied the "Pre-K Mathematics" curriculum, which included small group instruction and concrete manipulatives (i.e., blocks or other physical objects) for use in the classroom as well as software-based and home-based components. The authors randomized the intervention across 40 classrooms (20 Head Start, 20 state-funded) in California and New York. They also found large effects (0.57σ) on a childhood math assessment.

The two studies we uncovered that explore curricula that target social-emotional learning provided mixed results. Bierman et al. (2008) and Bierman et al. (2021) studied randomized controlled trials of a Research-based, Development Informed (REDI) curriculum in 44 Head Start classrooms on outcomes at ages 4-5 and on outcomes at ages 12-15 respectively. The intervention incorporates "brief lessons, 'hands-on' extension activities, and specific teaching strategies," designed to target social-emotional development and emergent literacy. Upshur et al. (2019) studied the Second Step Early Learning curriculum by randomizing its use across 67 classrooms in 13 sites. The intervention involves

Figure 8: Varying characteristics: curriculum



Notes: This figure plots effect sizes from studies that randomly vary a curriculum. The x-axis indicates the relevant study. The strip text denotes the outcome studies: behavior, emotion, math, and vocabulary/language. Estimates that are statistically significant at the 95 percent level are bold and opaque; statistically imprecise estimates or estimates from studies that did not report information sufficient to test for statistical significance are translucent. Colors denote distinct experiments. See appendix table A.8 for additional details on the construction of this figure, including the relevant point estimates.

"scripted, five-day-a-week, brief large and small group lessons with 28 weekly themes, along with suggested extension and generalization activities." Both studies appeared to generate small-to-moderate effects on some outcome categories but not others, which makes it hard to draw broad, generalizable lessons.

We uncovered two studies that found small to moderate positive effect sizes for the impact of the "Literacy Express" curriculum. This curriculum emphasizes the "use of specific teacher-initiated instructional activities that focus on key language and early literacy skills, a specific scope and sequence, and a significant use of small-group instructional activities" (Lonigan et al., 2011). Farver et al. (2009) studied its impact on Spanish-speaking English Language Learners. The authors randomized three classrooms containing 94 Spanish-speaking children into three treatment arms: a "business-as-usual" condition using the High/Scope curriculum[13] in English, an English-only Literacy Express Curriculum, or a transitional English version of the Literacy Express Curriculum. Lonigan et al. (2011) randomized 28 Head Start centers (739 children) into control groups that used the High/Scope curriculum and a treatment group that used Literacy Express.

Two important limitations span virtually all of these studies. The first is that the outcomes studied are often directly aligned to the curriculum under consideration. This makes it difficult to understand the full impact of the curriculum change, which may take time away from teaching other important skills. Another limitation is that the curricula studied in this section nearly always represent a bundled treatment. Elements such as small group instruction, the use of manipulatives, and the use of computers are never isolated from one another. That makes it difficult to synthesize results across studies and identify what program elements could be generalized to effectively teach a variety of skills across different domains.

4.2 Class size and length of school day

Evidence points to structural, non-pedagogical elements of the classroom environment that help drive quality. Atteberry et al. (2019) studied the effect of half-day versus full-day preschool in Colorado. The research design leveraged excessive demand for full-day spots. That allowed the research team to randomly assign offers to attend full-day preschool among the 226 applicants. The authors found that, at the end of preschool, the intervention generated large and statistically precise effects on measures of vocabulary

[13]The High/Scope curriculum features active learning, sharing decisions with children, a consistent daily routine, and ongoing assessment (HighScope, 2022).

skills (0.275σ), cognition (0.32σ), literacy (0.487σ), math (0.285σ), physical development (0.294σ), and social-emotional (0.19σ) development. These effect sizes are similar in magnitude to findings in the K-12 literature, such as the impact of randomizing children of a similar age to full-day kindergarten (Krueger, 2003; Gibbs, 2016). Although follow-up data was limited, the authors reported evidence that gains in literacy persisted after the children entered kindergarten the subsequent fall. That finding suggests that full-day preschool does achieve its stated goal of increasing kindergarten readiness.

Francis and Barnett (2019) studied the impact of capping class sizes in Chicago. They recruited 22 teachers, who each taught two sections (one morning, one afternoon, each lasting 2.5 hours) for a total of 44 sessions. The study collected test score data from the 354 children who were enrolled. The research design involved randomly choosing a morning or afternoon session for each teacher and capping it at 15 students. The default class size was 20. Five teachers did not comply with the random assignment. The authors explored three test score-based measures of student learning and found statistically precise evidence of literacy gains in the capped classes (0.2σ). However, given the small sample sizes involved, the problems with statistical inference noted at the beginning of this section, and the fact that the authors did not adjust for the non-compliance, the findings from this study should be taken as suggestive at best.

4.3 Language / immersion

Immersing Spanish-speaking children in a Spanish language classroom appears to enhance their Spanish language vocabulary and acquisition, with little negative impact on their English skills. However, there do not appear to be large benefits across other measures of cognition and learning. These conclusions are supported by consistent evidence from three studies that randomly varied the foreign language content of classrooms.

Barnett et al. (2007) randomized nearly 1,000 students in a northeastern city who applied to an oversubscribed two-way language (Spanish) immersion program. Children who did not win the lottery were offered spots in a standard English language classroom. The study team recruited 147 children, a mixture of native and non-native English speakers, to collect follow-up outcomes. The two-way immersion program generated gains in Spanish language vocabulary among Spanish-speaking students.

Durán et al. (2010) studied the impact of assignment to a transitional bilingual education model. Unlike the intervention studied in Barnett et al. (2007), the goal of the treatment studied in Durán et al. (2010) related specifically to supporting non-native

English-speaking students as part of a transition to English language instruction. The study team recruited 31 Spanish-speaking children aged 3-4 and randomized half to the transitional bilingual education model and the other half to an English-only classroom. Consistent with the results from Barnett et al. (2007), the research team found that children in the treated group scored higher on measures of Spanish language vocabulary and aptitude at endline with no detectable impact (positive or negative) on English language ability/acquisition.

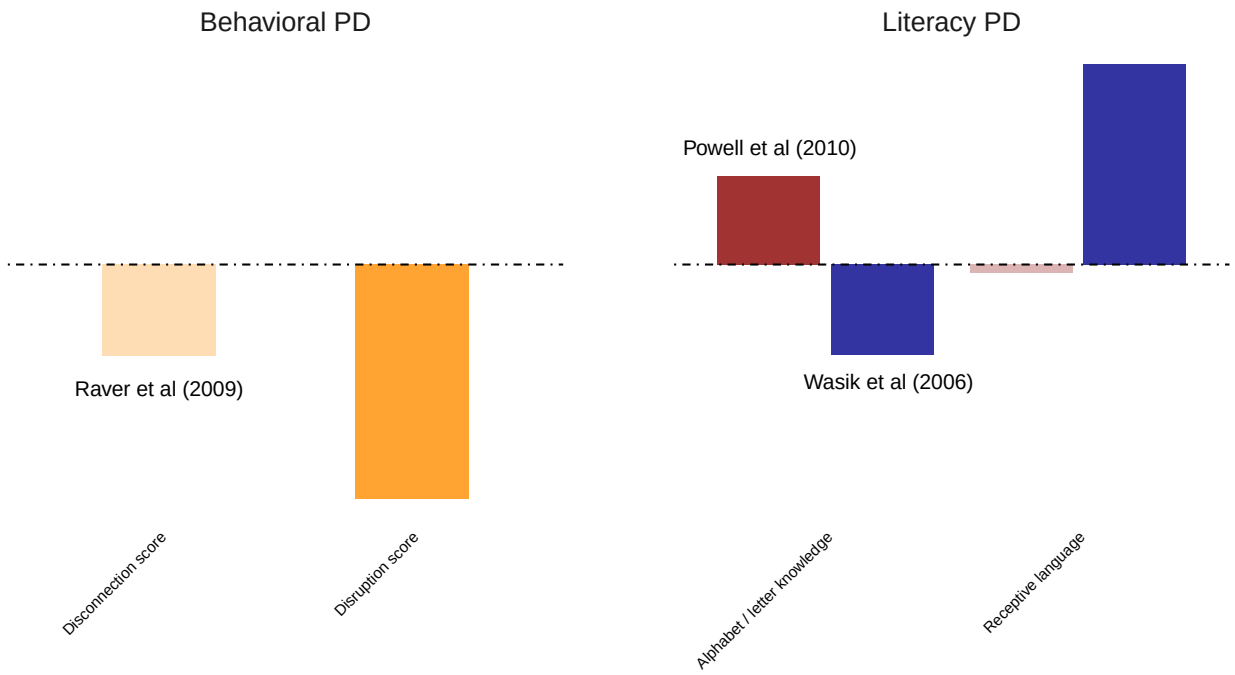
Finally, we return to the Farver et al. (2009) study discussed in section 4.1. It randomly varied an early literacy curriculum across three classrooms, with a total of 94 Spanish-speaking students. As with the preceding two studies, the researchers found large effect sizes ($0.48 - 0.83\sigma$) on measures of Spanish language acquisition and vocabulary. Unlike the other two studies, they also found moderate positive effect sizes on English language acquisition compared with their "business-as-usual" control group; however, they did not find differences in English acquisition relative to the English-only treatment arm, suggesting that the broader benefits outside Spanish language acquisition emerged from the curriculum itself rather than the Spanish immersion component.

4.4 Professional development

The evidence is mixed on the importance of teacher professional development. Our search uncovered three studies that randomly varied professional development. Wasik et al. (2006) examined a training program about strategies for teaching book reading and oral language. They randomized the intervention across 16 teachers spanning two Head Start sites covering 207 children aged 2 years, 8 months to 4 years, 10 months. Raver et al. (2009) studied a development program for addressing behavioral problems. It contained four components: (1) teacher training; (2) ongoing coaching; (3) stress reduction workshops; and (4) targeted services to select children with the highest social-emotional needs. The researchers matched 18 Head Start sites on the basis of observable characteristics and randomized the intervention to nine of them. Within sites, a total of 35 classrooms participated covering more than 500 children. Powell et al. (2010) explored an intervention that offered teachers a two-day workshop on techniques for improving English literacy skills followed by expert coaching. The random assignment occurred at the teacher level and involved a total of 759 students enrolled in 88 classrooms spanning 24 centers.

Figure 9 plots effect sizes from these studies. The x-axis indicates test score-based measures of the indicated concept. The left-hand panel shows results from the Raver et

Figure 9: Varying characteristics: professional development



Notes: This figure plots effect sizes from Wasik et al. (2006), Raver et al. (2009), and Powell et al. (2010). The x-axis indicates test score-based measures of the indicated concept. The left-hand side panel shows results from the Raver et al. (2009) study of the training program involving behavior. The right-hand panel shows results from the two studies focused on literacy programs. Estimates that are statistically significant at the 95 percent level are bold and opaque; statistically imprecise estimates are translucent. See appendix table A.9 for additional details on the construction of this figure, including the relevant point estimates.

al. (2009) study of the behavioral program. The right-hand panel shows results from the two studies that focused on literacy programs. Estimates statistically significant at the 95 percent level are bold and opaque; statistically imprecise estimates are translucent.

The results in figure 9 reveal a mix of effects, with no consistent patterns emerging. The effect sizes from Raver et al. (2009) are moderate to large and negative (-$0.33 - 0.85\sigma$), suggesting the intervention may have affected the intended outcome by causing a decline in behavioral issues; however, the effect of the intervention is significant only for one of the two outcomes explored. At the same time, the early literacy interventions are more mixed. Of the four point estimates, two are statistically significant and move in the positive direction, one is statistically insignificant, and a third is negative. As a result, it is difficult to draw any clear message from these findings, especially in light of the

small effective sample sizes involved in the randomization and the concerns raised at the beginning of this section about statistical inference in this literature.[14]

5 Critical knowledge gaps

While researchers have made much progress on the subject of preschool in the United States, key knowledge gaps remain. In this section, we highlight three that we believe are most critical.

5.1 Learning the causal drivers of effective preschool

As highlighted in section 4, only a few studies exploit random variation to uncover the characteristics of effective preschool. This literature tends to be limited by small samples and non-robust approaches to statistical inference. As a result, little definitive evidence is available. Yet, this is arguably the most critical question facing early childhood education policymakers today. Given the dominant position of the government in the early childhood education sector, standards and regulations regarding state and federal funding carry the potential to reshape preschool at large. More information about the drivers of preschool quality would benefit not only the students brought into the system by expansions but also the many students who would have enrolled anyway.

Which aspects of preschool should be prioritized for study? The length of school day appears to be a key factor, given the results in Atteberry et al. (2019), the corroborating heterogeneity in Walters (2015), and similar patterns in kindergarten documented in Gibbs (2016). In light of promising indications that curriculum influences academic achievement, additional work exploring the constituent elements of curricula design (e.g., small group instruction and computer-based learning) potentially could translate into multiple skill domains.

Understanding what makes an effective early childhood teacher also warrants further study. We could find virtually no experimental work on this topic outside the limited evidence on professional development highlighted in section 4.4. Further research would help policymakers and site administrators make informed decisions regarding hiring and

[14]The effective N in these studies ranges from 18 to 88, since the randomization typically occurred at the teacher/site/classroom level. However, it is unclear from the published papers whether the inferential procedure employed by the authors clustered in a way that would properly account for the dependence among students assigned to the same classroom when hypothesis testing.

firing, qualifications, and compensation. These decisions are particularly salient in light of the shortage of early childhood educators (Coffey and Khattar, 2022). Another critical area for further research is how to measure classroom effectiveness, such as through lottery-based studies. By contrast, the K-12 system can draw on extensive literature on the role of test scores and in-person observations.

5.2 Developing alternative short-term outcome measures

A fundamental challenge of studying the impact of preschool is that the most consequential outcomes often cannot be measured until at least a decade has passed. For that reason, researchers have typically relied on cognitive test scores to evaluate intervention efficacy. However, as the fade-out patterns from both the demonstration studies and the modern work make clear, the link between cognitive test scores and the distal outcomes is not straightforward. This limitation is part of a broader set of critiques of the modern preschool assessment systems. For example, scholars have noted that the existing assessment systems were developed with samples that are not representative of the children actually attending preschool in the United States (MDRC, 2023). These same critics also argue that assessment systems are expensive to administer and frequently fail to provide useful insight to families and teachers (MDRC, 2023).

Alternative short-run outcome measures are needed to better capture the channels connected to long-term success and well-being, in particular the non-cognitive skills believed by researchers to be a key mechanism. For example, future longitudinal studies that collect original data should include tests, such as the Challenging Situations Task, designed explicitly to capture non-cognitive skills (Denham et al., 2014). Researchers should also develop novel ways of measuring these concepts based on real-world behavior rather than test responses. For example, economists have amassed a rich, laboratory-based literature that uses incentivized games to isolate key economic parameters through small- to medium-stakes choices rather than relying on self-reports (Kagel and Roth, 2020). A similar approach could be productive here.

Scholars should also develop new ways of leveraging existing data to measure relevant concepts such as non-cognitive skills and grit. Much empirical work relies on pre-existing administrative data collected for government or industry purposes. This makes it infeasible to field survey instruments to the study subjects. One intriguing approach involves re-purposing existing item response data from state standardized tests to capture patterns of student answers that better connect to long-run outcomes and/or other

interesting concepts (Bruhn et al., 2023). Alternative approaches could involve validating outcomes–ideally long-run–commonly found in school district administrative data against standard non-cognitive test batteries.

Finally, we note that economic theory could provide valuable guidance. While social psychologists may have a precise definition of the term "non-cognitive skill," we believe that this term frequently functions as a catch-all for "everything that isn't captured by test scores." Yet, we know very little about the actual mechanisms within the "non-test score bucket" that drive long-run persistence. It could be that preschool indeed teaches students essential skills like focus, curiosity, and grit/resilience. Alternatively, preschool may catalyze subsequent parental investments, as documented in Gelber and Isen (2013). Understanding which mechanism is at work has strong implications both for the design of effective early childhood interventions *and* for the types of measures that could gauge efficacy in the short run. However, disentangling competing mechanisms requires strong, testable theoretical predictions to guide the data researchers collect and the empirical exercises they perform. We are aware of little work on this topic.

5.3 Heterogeneity, equity, and community impact

Parental labor supply and the broader community. Little lottery-based work has been conducted on the impact of preschool beyond the direct effects on children. Future research should explore the broader impacts on the family, particularly parents' employment status. As pointed out in Cascio (2015, 2021), preschool serves both educational and child-care purposes. If access to preschool, especially subsidized slots for low-income families, frees parents to work or accept better-paying jobs, that could significantly affect cost-benefit analysis for state funding. As far as we know, the only experimental evidence on this question comes from the conflicting conclusions of three papers that used random assignment data from the Head Start impact study (Sabol and Chase-Lansdale, 2015; Schiman, 2022; Wikle and Wilson, 2023). The limited US-based quasi-experimental work on this topic also produced mixed results. (Fitzpatrick, 2010, 2012; Pihl, 2018).

Much research has been conducted in the K-12 system on equilibrium/spillover effects and the broader impact of educational policy changes on educator labor markets (e.g. Allende, 2019; Bobba et al., 2021; Bruhn et al., 2022; Campos and Kearns, 2022, to name a few recent examples), but virtually none has been done in the preschool setting. Leveraging lotteries to explore such wide-ranging topics is essential for a rich, nuanced understanding of the overall impact of the early childhood education system.

Equity, geography, and sub-group diversity The current body of work on US preschools is heavily skewed in its geographic and demographic representation. The ≈ 350 children involved in the early demonstration programs came from extremely low-income families that lived in rural areas and were almost exclusively Black (see tables 2 and 1). Modern studies also tend to focus on lower-income children from disproportionately Black families (see tables 4 and 3). This is because modern estimates come from programs that are either means-tested (as in the Headstart Impact Study and Educare) or that happen to serve urban or lower socio-economic status populations (as in Boston and Tennessee).

Since the existing body of work is so concentrated on non-representative samples, it is less likely to generalize to the populations being considered as targets for expansion. This is especially consequential in light of the current push for preschool expansion as an engine for equity (McSorley, 2023). Additional work using lottery-based research designs with study subjects encompassing a broader set of geographic and demographic sub-groups is necessary for deciding where to expand publicly funded preschool and which sub-populations to target.

6 Conclusion: opportunities for future work

The lottery evidence on preschool yields a number of consistent patterns. Large gains on cognitive test scores quickly fade, only to reemerge in consequential later-life outcomes related to well-being. Non-cognitive skills appear to be an important channel. Differences in effect sizes over time can be explained by increased preschool enrollment in the control group in modern studies and the unusual quality/intensity of the earliest demonstration studies. Preschool sites vary widely in effectiveness, suggesting that identifying and implementing effective practices would boost overall performance within the system. That would benefit both current students and those drawn in by expansions.

Yet critical gaps remain. Very little lottery evidence addresses the policies and practices that boost quality, and existing evidence suffers from methodological issues related to statistical inference that can affect the reliability of the conclusions. Further, we know very little about the concrete non-cognitive channels responsible for transmitting preschool experiences into adult well-being. Given the consistent pattern of fade-out on cognitive test scores, this raises important questions about appropriate ways to measure short-run effectiveness of preschool programs.

Currently, the United States appears poised to enter an era of unprecedented preschool

expansion (Potts, 2023). This presents a generational opportunity for policymakers to partner with researchers. As seen in the work from the modern era, simply facilitating access to existing data, such as the application files from an oversubscribed preschool program, can yield valuable insights for policy design. More ambitious approaches could leverage the latitude policymakers often have in rolling out interventions and the design and timing of rules and regulations. By randomly varying implementation among different subsets of the relevant sector, researchers could better learn what works and doesn't work.

Ideally, collaborations between academics and policymakers will lead to effective research-practice partnerships (Wentworth et al., 2023). Establishing long-running relationships has the potential to generate actionable, ongoing, data-driven insights into the day-to-day operation of a preschool system. At the same time, academics can often find broader lessons in the specific questions that arise day-to-day. In turn, those lessons could be applied widely to improve the early childhood education system as a whole.

References

Abdulkadiroğlu, Atila, Joshua D Angrist, Yusuke Narita, and Parag A Pathak, "Research design meets market design: Using centralized assignment for impact evaluation," *Econometrica*, 2017, *85* (5), 1373–1432.

Allende, Claudia, "Competition under social interactions and the design of education policies," *Job Market Paper*, 2019.

Anderson, Michael L, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495.

Angrist, Joshua D and Jörn-Steffen Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press, 2009.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang, "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely," Technical Report, National Bureau of Economic Research 2019.

Atteberry, Allison, Daphna Bassok, and Vivian C Wong, "The effects of full-day prekindergarten: Experimental evidence of impacts on children's school readiness," *Educational Evaluation and Policy Analysis*, 2019, *41* (4), 537–562.

Barnett, W Steven, Donald J Yarosz, Jessica Thomas, Kwanghee Jung, and Dulce Blanco, "Two-way and monolingual English immersion in preschool education: An experimental comparison," *Early Childhood Research Quarterly*, 2007, *22* (3), 277–293.

Bierman, Karen L, Brenda S Heinrichs, Janet A Welsh, and Robert L Nix, "Reducing adolescent psychopathology in socioeconomically disadvantaged children with a preschool intervention: A randomized controlled trial," *American Journal of Psychiatry*, 2021, *178* (4), 305–312.

_ , Celene E Domitrovich, Robert L Nix, Scott D Gest, Janet A Welsh, Mark T Greenberg, Clancy Blair, Keith E Nelson, and Sukhdeep Gill, "Promoting academic and social-emotional school readiness: The Head Start REDI program," *Child Development*, 2008, *79* (6), 1802–1817.

Bobba, Matteo, Tim Ederer, Gianmarco Leon-Ciliotta, Christopher Neilson, and Marco G Nieddu, "Teacher compensation and structural inequality: Evidence from centralized teacher school choice in Perú," Technical Report, National Bureau of Economic Research 2021.

Bowles, Samuel and Herbert Gintis, "Schooling in capitalist America. New York: Basic book," *Inc. Publishers*, 1976.

_ and _ , "The inheritance of economic status: Education, class and genetics," *International Encyclopedia of the Social and Behavioral Sciences: Genetics, Behavior and Society*, 2001, *6*, 4132–141.

Brey, Cristobal De, Thomas D Snyder, Anlan Zhang, and Sally A Dillow, "Digest of Education Statistics 2019. NCES 2021-009.," *National Center for Education Statistics*, 2021.

Bruhn, Jesse, Michael Gilraine, Jens Ludwig, and Sendhil Mullainathan, "What's in a Question? Using item response data to better represent learning.," *Unpublished manuscript*, 2023.

_ , **Scott Imberman, and Marcus Winters**, "Regulatory arbitrage in teacher hiring and retention: Evidence from Massachusetts charter schools," *Journal of Public Economics*, 2022, *215*, 104750.

Campbell, Frances A and Craig T Ramey, "Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention," *American Educational Research Journal*, 1995, *32* (4), 743–772.

_ , **Barbara H Wasik, Elizabeth Pungello, Margaret Burchinal, Oscar Barbarin, Kirsten Kainz, Joseph J Sparling, and Craig T Ramey**, "Young adult outcomes of the Abecedarian and CARE early childhood educational interventions," *Early Childhood Research Quarterly*, 2008, *23* (4), 452–466.

Campos, Christopher and Caitlin Kearns, "The impact of neighborhood school choice: Evidence from Los Angeles' zones of choice," in "The Impact of Neighborhood School Choice: Evidence from Los Angeles' Zones of Choice: Campos, Christopher— Kearns, Caitlin," [Sl]: SSRN, 2022.

Cascio, Elizabeth U, "The promises and pitfalls of universal early education," *IZA World of Labor*, 2015.

_ , "Early childhood education in the United States: What, when, where, who, how, and why," Technical Report, National Bureau of Economic Research 2021.

_ **and Douglas O Staiger**, "Knowledge, tests, and fadeout in educational interventions," Technical Report, National Bureau of Economic Research 2012.

Chabrier, Julia, Sarah Cohodes, and Philip Oreopoulos, "What can we learn from charter school lotteries?," *Journal of Economic Perspectives*, 2016, *30* (3), 57–84.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review*, 2014, *104* (9), 2633–2679.

_ , _ , **Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan**, "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

Clements, Douglas H, Julie Sarama, Mary Elaine Spitler, Alissa A Lange, and Christopher B Wolfe, "Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial," *Journal for Research in Mathematics Education*, 2011, *42* (2), 127–166.

Coffey, Maureen and Rose Khattar, "The child care sector will continue to struggle hiring staff unless it creates good jobs," *Center for American Progress Action Fund*, 2022.

Conti, Gabriella, James J Heckman, and Rodrigo Pinto, "The effects of two influential early childhood interventions on health and healthy behaviour," *The Economic Journal*, 2016, *126* (596), F28–F65.

Denham, Susanne A, Hideko Hamada Bassett, Erin Way, Sara Kalb, Heather Warren-Khot, and Katherine Zinsser, ""How would you feel? What would you do?" Development and underpinnings of preschoolers' social information processing," *Journal of Research in Childhood Education*, 2014, *28* (2), 182–202.

Duncan, Greg, Ariel Kalil, Magne Mogstad, and Mari Rege, "Investing in early childhood development in preschool and at home," 2022.

Duncan, Greg J and Katherine Magnuson, "Investing in preschool programs," *Journal of economic perspectives*, 2013, *27* (2), 109–132.

Durán, Lillian K, Cary J Roseth, and Patricia Hoffman, "An experimental study comparing English-only and transitional bilingual education on Spanish-speaking preschoolers' early literacy development," *Early Childhood Research Quarterly*, 2010, *25* (2), 207–217.

Durkin, Kelley, Mark W Lipsey, Dale C Farran, and Sarah E Wiesen, "Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade.," *Developmental Psychology*, 2022, *58* (3), 470.

Farver, Jo Ann M, Christopher J Lonigan, and Stefanie Eppe, "Effective early literacy skill development for young Spanish-speaking English language learners: An experimental study of two methods," *Child Development*, 2009, *80* (3), 703–719.

Feller, Avi, Todd Grindal, Luke Miratrix, and Lindsay C Page, "Compared to what? Variation in the impacts of early childhood education by alternative care type," 2016.

Fitzpatrick, Maria Donovan, "Preschoolers enrolled and mothers at work? The effects of universal prekindergarten," *Journal of Labor Economics*, 2010, *28* (1), 51–85.

_ , "Revising our thinking about the relationship between maternal labor supply and preschool," *Journal of Human Resources*, 2012, *47* (3), 583–612.

Francis, Jessica and William Steven Barnett, "Relating preschool class size to classroom quality and student achievement," *Early Childhood Research Quarterly*, 2019, *49*, 49–58.

Friedman-Krauss, Allison, Barnett W Steven, A Garver Karin, S Hodges Katherine, GG Weisenfeld, and Gardiner Beth Ann, "The State of Preschool: 2019 State Preschool Yearbook," *National Institute for Early Education Research Report, Rutgers University*, 2019.

García, Jorge Luis, James J Heckman, and Victor Ronda, "The lasting effects of early-Childhood education on promoting the skills and social mobility of disadvantaged African Americans and their children," *Journal of Political Economy*, 2023, *131* (6), 000–000.

Gelber, Alexander and Adam Isen, "Children's schooling and parents' behavior: Evidence from the Head Start Impact Study," *Journal of Public Economics*, 2013, *101*, 25–38.

Gelman, Andrew and Jennifer Hill, *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press, 2006.

Gibbs, Chloe R, "Treatments, peers, and treatment effects in full-day kindergarten: Reconciling experimental and quasi-experimental impact evidence," *Manuscript]. https://pdfs. semanticscholar. org/160e/a5248d6902bd1e11afea8af060c0ddd05bb7. pdf*, 2016.

Gray-Lobe, Guthrie, Parag A Pathak, and Christopher R Walters, "The long-term effects of universal preschool in Boston," *The Quarterly Journal of Economics*, 2023, *138* (1), 363–411.

Hansen, Bruce, *Econometrics*, Princeton University Press, 2022.

Heckman, James J, Jora Stixrud, and Sergio Urzua, "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor Economics*, 2006, *24* (3), 411–482.

_ , **Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev, and Adam Yavitz**, "The rate of return to the HighScope Perry Preschool Program," *Journal of Public Economics*, 2010, *94* (1-2), 114–128.

Heckman, James, Rodrigo Pinto, and Peter Savelyev, "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes," *American Economic Review*, 2013, *103* (6), 2052–2086.

_ , **Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz**, "Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program," *Quantitative Economics*, 2010, *1* (1), 1–46.

HighScope, Jan 2022.

Jackson, C Kirabo and Claire L Mackevicius, "What impacts can we expect from school spending policy? Evidence from evaluations in the US," *American Economic Journal: Applied Economics*, 2023.

Jr, Joseph H Stevens, "From 3 to 20: The early training project," *Young Children*, 1982, pp. 57–64.

Kagel, John H and Alvin E Roth, *The handbook of experimental economics, volume 2*, Princeton University Press, 2020.

Klaus, Rupert A and Susan W Gray, "The early training project for disadvantaged children: A report after five years," *Monographs of the Society for Research in Child Development*, 1968, *33* (4), iii–66.

Klein, Alice, Prentice Starkey, Douglas Clements, Julie Sarama, and Roopa Iyer, "Effects of a pre-kindergarten mathematics intervention: A randomized experiment," *Journal of Research on Educational Effectiveness*, 2008, *1* (3), 155–178.

Kline, Patrick and Christopher R Walters, "Evaluating public programs with close substitutes: The case of Head Start," *The Quarterly Journal of Economics*, 2016, *131* (4), 1795–1848.

Krueger, Alan B, "Economic considerations and class size," *The Economic Journal*, 2003, *113* (485), F34–F63.

Lipsey, Mark W, Dale C Farran, and Kelley Durkin, "Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade," *Early Childhood Research Quarterly*, 2018, *45*, 155–176.

Lonigan, Christopher J, JoAnn M Farver, Beth M Phillips, and Jeanine Clancy-Menchetti, "Promoting the development of preschool children's emergent literacy skills: A randomized evaluation of a literacy-focused curriculum and two professional development models," *Reading and Writing*, 2011, *24*, 305–337.

McCarton, Cecelia M, Jeanne Brooks-Gunn, Ina F Wallace, Charles R Bauer, Forrest C Bennett, Judy C Bernbaum, R Sue Broyles, Patrick H Casey, Marie C McCormick, David T Scott et al., "Results at age 8 years of early intervention for low-birth-weight premature infants: The Infant Health and Development Program," *Jama*, 1997, *277* (2), 126–132.

McCoy, Dana Charles, Hirokazu Yoshikawa, Kathleen M Ziol-Guest, Greg J Duncan, Holly S Schindler, Katherine Magnuson, Rui Yang, Andrew Koepp, and Jack P Shonkoff, "Impacts of early childhood education on medium-and long-term educational outcomes," *Educational Researcher*, 2017, *46* (8), 474–487.

McSorley, Laura Dallas, "6 ways to ensure preschool contributes to an equitable early childhood system," Oct 2023.

MDRC, "Measures for early success," Aug 2023.

Obama, Barack, "Remarks by the president in Working Mothers Town Hall," 2015. Remarks by President Barack Obama in Charlotte, N.C.

Pihl, Ariel Marek, "Head Start and mothers' Work: Free child care or something more?," Technical Report 2018.

Popli, Nick and Abby Vesoulis, "The house just passed Biden's build back better bill. Here's what's in it.," *Time*, November 2021.

Potts, Monica, "Why more states don't have universal pre-K?," *FiveThirtyEight*, February 2023.

Powell, Douglas R, Karen E Diamond, Margaret R Burchinal, and Matthew J Koehler, "Effects of an early literacy professional development intervention on Head Start teachers and children," *Journal of Educational Psychology*, 2010, *102* (2), 299.

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, Gary Shapiro, Pam Broene, Frank Jenkins, Philip Fletcher, Liz Quinn, Janet Friedman et al., "Head Start Impact Study. Final Report.," *Administration for Children & Families*, 2010.

Pungello, Elizabeth P, Kirsten Kainz, Margaret Burchinal, Barbara H Wasik, Joseph J Sparling, Craig T Ramey, and Frances A Campbell, "Early educational intervention, early cumulative risk, and the early home environment as predictors of young adult outcomes within a high-risk sample," *Child Development*, 2010, *81* (1), 410–426.

Raver, C Cybele, Stephanie M Jones, Christine Li-Grining, Fuhua Zhai, Molly W Metzger, and Bonnie Solomon, "Targeting children's behavior problems in preschool classrooms: a cluster-randomized controlled trial," *Journal of Consulting and Clinical Psychology*, 2009, *77* (2), 302.

Reynolds, Arthur J, Judy A Temple, Suh-Ruu Ou, Irma A Arteaga, and Barry AB White, "School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups," *Science*, 2011, *333* (6040), 360–364.

Ricciuti, Anne E, Robert G St Pierre, Wang Lee, and Amanda Parsad, "Third National Even Start Evaluation: Follow-Up findings from the experimental design study. NCEE 2005-3002.," *National Center for Education Evaluation and Regional Assistance NCEE*, 2004.

Sabol, Terri J and P Lindsay Chase-Lansdale, "The influence of low-income children's participation in Head Start on their parents' education and employment," *Journal of Policy Analysis and Management*, 2015, *34* (1), 136–161.

Schiman, Cuiping, "Experimental evidence of the effect of Head Start on mothers' labor supply and human capital investments," *Review of Economics of the Household*, 2022, *20* (1), 199–241.

Schweinhart, Lawrence J, *The High/Scope Perry Preschool study through age 40: Summary, conclusions, and frequently asked questions*, High/Scope Educational Research Foundation, 2004.

_ **, Howard V Barnes, and David P Weikhart**, "Significant benefits: The High/Scope Perry preschool study through age 27," *Child welfare: Major themes in health and social welfare*, 2005, *4*, 9–29.

Upshur, Carole C, Melodie Wenz-Gross, Christopher Rhoads, Miriam Heyman, Yeon-soo Yoo, and Gail Sawosik, "A randomized efficacy trial of the second step early learning (SSEL) curriculum," *Journal of Applied Developmental Psychology*, 2019, *62*, 145–159.

US Department of Education, 2019.

US Department of Health and Human Services, "Head Start Program Facts: Fiscal Year 2021," https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/hs-program-fact-sheet-2021.pdf. Accessed: 2023-09-13.
U.S. Code Title 45 - Public Welfare

U.S. Code Title 45 - Public Welfare.

Walters, Christopher R, *"Inputs in the production of early childhood human capital: Evidence from Head Start," American Economic Journal: Applied Economics, 2015, 7 (4), 76–102.*

Wasik, Barbara A, Mary Alice Bond, and Annemarie Hindman, *"The effects of a language and literacy intervention on Head Start children and teachers," Journal of Educational Psychology, 2006, 98 (1), 63.*

Weiland, Christina, Rebecca Unterman, Anna Shapiro, Sara Staszak, Shana Rochester, and Eleanor Martin, *"The effects of enrolling in oversubscribed prekindergarten programs through third grade," Child Development, 2020, 91 (5), 1401–1422.*

Wentworth, Laura, Paula Arce-Trigatti, Carrie Conaway, and Samantha Shewchuk, *Brokering in education research-practice partnerships: A guide for education professionals and researchers, Taylor & Francis, 2023.*

Wikle, Jocelyn and Riley Wilson, *"Access to Head Start and maternal labor supply: experimental and quasi-experimental evidence," Journal of Labor Economics, 2023, 41 (4), 000–000.*

Yazejian, Noreen, Donna M Bryant, Laura J Kuhn, Margaret Burchinal, Diane Horm, Sydney Hans, Nancy File, and Barbara Jackson, *"The Educare intervention: Outcomes at age 3," Early Childhood Research Quarterly, 2020, 53, 425–440.*

_ , _ , Sydney Hans, Diane Horm, Lisa St. Clair, Nancy File, and Margaret Burchinal, *"Child and parenting outcomes after 1 year of Educare," Child Development, 2017, 88 (5), 1671–1688.*

A Additional detail regarding figure construction

Table A.1: Additional details for figure 1

intervention	outcome_group	effect_size	standard_error
ABC	Kindergarten (Age 5)	0.604	0.24
PPP	Kindergarten (Age 5)	0.871	0.214
ETP	Kindergarten (Age 5)	0.51	0.198
Pooled	Kindergarten (Age 5)	0.689	0.117
ABC	High school (Ages 14-17)	0.432	0.244
PPP	High school (Ages 14-17)	0.176	0.177
ETP	High school (Ages 14-17)	0.136	0.243
Pooled	High school (Ages 14-17)	0.269	0.121

Notes: This table provides the underlying point estimates and standard errors used to construct figure 1 in the main text. Tables 1 and 4 in Anderson (2008) provide the source point estimates, outcome standard deviations, and standard errors used to create the effect sizes in this table. In this table, program-specific estimates are constructed by first standardizing point estimates from Anderson (2008) by the corresponding outcome standard deviation and then taking a precision-weighted average across genders. Pooled estimates represent precision-weighted averages across demonstration programs.

Table A.2: Additional details for figure 2

outcome_variable	effect_size	standard_error
Graduated HS	0.168	0.094
Education after highschool	-0.103	0.179
Employed (age 21-27)	0.059	0.092
Annual Income (age 27)	0.222	0.175
Annual Income (age 40)	0.253	0.21
Teen parent	0.313	0.139
Criminal record	0.14	0.128
Married	0.326	0.4
Drug use	0.447	0.182
Index	0.154	0.066

Notes: This table provides the underlying point estimates and standard errors used to construct figure 2 in the main text. Anderson (2008) provides the source point estimates, outcome means, and standard errors used to create the average effect sizes in this table. Specifically, table 3 provides the index of adult outcomes (ABC, PPP, and ETP); table 6 provides the high school graduation outcome (ABC, PPP and, in the case of ETP, ever dropped out of high school); table 7 provides the teen parent outcome (ABC and PPP); table 8 provides education after high school (in-college at 21 for ABC, any-college at 27 for PPP, and in-post high-school education at 21 for ETC); table 9 provides the employed at age 21-27 outcomes (ABC at age 21 and PPP at age 27) and annual income at ages 27 and 40 (PPP only); table 10 provides the criminal record outcome (PPP only), drug use (marijuana for ABC and any drug use for PPP), and marriage (PPP only). Pooled estimates in this table are constructed by dividing point estimate and standard errors by the corresponding outcome mean and then taking precision-weighted averages across genders and demonstration programs.

Table A.3: Additional details for figure 3

intervention	outcome_age	effect_size	standard_error
BOS	8-9	0.024	0.094
BOS	9-10	-0.063	0.066
BOS	10-11	0.022	0.076
BOS	11-12	-0.023	0.067
BOS	12-13	-0.003	0.064
BOS	13-14	0.024	0.063
BOS	15-16	-0.031	0.064
Educare	3-4	0.288	0.131
HSIS	4-5	0.247	0.031
HSIS	5-6	0.093	0.049
HSIS	6-7	0.049	0.05
Tennessee	4-5	0.395	0.102
Tennessee	5-6	0.032	0.077
Tennessee	6-7	-0.024	0.085
Tennessee	7-8	-0.132	0.087
Tennessee	8-9	-0.115	0.091
Tennessee	11-12	-0.333	0.17

Notes: This table provides the underlying point estimates and standard errors used to construct figure 3 in the main text. Estimates from Boston represent impacts on state standardized math tests and are taken from table IV in Gray-Lobe et al. (2023). These estimates were already converted to effect sizes. Grade-level specific estimates are mapped to the typical age range of children appearing in that grade. The point estimate for Educare is the ITT estimate for the WJ-3 AP outcome contained table 5 of Yazejian et al. (2020) and is scaled by the corresponding control group standard deviation in table 2. Estimates for the Headstart Impact Study represent impacts on an average of WJ-3 and PPVT scores and are taken from the pooled IV estimates column of table 2 in Kline and Walters (2016). These estimates were already converted to effect sizes. Age ranges are estimated using the starting age range of children in the study and the "time relative to treatment" column contained in the table 2. The TOT column from table 7 in Lipsey et al. (2018) provide impacts on the WJ-3 composite for Tennessee in pre-K through third grade. Grade-level specific estimates are mapped to the typical age range of children appearing in that grade. These estimates were already converted to effect sizes. Standard errors were not reported in this study and are instead bounded using reported levels of statistical significance and assuming a normal sampling distribution. The unweighted TOT estimates from table 2 in Durkin et al. (2022) provide impacts on state level standardized math tests for Tennessee in grades three and six. Grade-level specific estimates are mapped to the typical age range of children appearing in that grade. These estimates were already converted to effect sizes. Standard errors were not reported in this study and are instead reverse engineered from p-values assuming a normal sampling distribution.

Table A.4: Additional details for figure 4

intervention	outcome_var	outcome_grade	effect_size	standard_error
Boston	Days absent	6-8	0.012	0.069
Boston	Days absent	9-12	-0.098	0.076
Boston	Diverse learner	Grades: K-3	0.051	0.351
Tennessee	Preparedness K	Grades: K-3	0.056	0.032
Tennessee	Preparedness G3	Grades: K-3	-0.021	0.038
Tennessee	Retained in grade	Grades: K-3	0.007	0.243
Tennessee	Diverse learner	6-8	0.738	0.287
Tennessee	Above grade level	6-8	-0.018	0.029

Notes: This table provides the underlying point estimates and standard errors used to construct figure 4 in the main text. In all cases, effect sizes in this table are constructed by dividing source estimates by the associated outcome mean. Source estimates, standard errors, and control group outcome means used to create effects sizes for days absent in Boston are drawn from the 2SLS estimates from table VII in Gray-Lobe et al. (2023). Source estimates, complier outcome means, and standard errors used to create effects sizes for the "diverse learner" outcome in Boston are drawn from the "ever sped" outcome and CACE column of table 3 in Weiland et al. (2020). Source estimates and control means for preparedness and retained in grade (authors' preferred multiple imputation estimate) in Tennessee are taken from tables eight and 11 in Lipsey et al. (2018). Standard errors were not reported in this study and are instead reverse engineered from p-values assuming a normal sampling distribution. Source estimates, control means, and standard errors for the "diverse learner" and above grade level outcomes in Tennessee are taken from the IEP and on-grade rows (observed value) in table 4 of Durkin et al. (2022). Standard errors were not reported in this study and are instead reverse engineered from p-values assuming a normal sampling distribution.

Table A.5: Additional details for figure 5

intervention	outcome_var	outcome_grade	effect_size	standard_error
Boston	Suspensions	6-8	-0.221	0.253
Boston	Suspensions	9-12	-0.363	0.213
Boston	Incarceration	6-8	-1	1
Boston	Incarceration	9-12	-1.429	0.714
Tennessee	All offenses	Grades: K-3	0.239	0.286
Tennessee	All offenses	6-8	0.312	0.139
Tennessee	Major offenses	Grades: K-3	-0.029	0.38
Tennessee	Major offenses	6-8	0.477	0.236

Notes: This table provides the underlying point estimates and standard errors used to construct figure 5 in the main text. In all cases, effect sizes in this table are constructed by dividing source estimates by the associated outcome mean. Source estimates, standard errors, and control group outcome means used to create effects sizes for suspensions and incarceration in Boston are drawn from the 2SLS estimates from table VII in Gray-Lobe et al. (2023). Source estimates and control group outcome means for the K-3 outcomes in Tennessee are taken from the "observed values" row of table 13 in Lipsey et al. (2018). Standard errors were not reported in this study and are instead reverse engineered from p-values assuming a normal sampling distribution. Source estimates and control group outcome means for the grades 6-8 outcomes in Tennessee are taken from the TOT column in table five of Durkin et al. (2022). Standard errors were not reported in this study and are instead reverse engineered from p-values assuming a normal sampling distribution.

Table A.6: Additional details for figure 6

outcome_var	ages	effect_size	standard_error
All	During pre-k elligibility (ages 3-4)	0.15	0.02
All	After pre-k elligibility (ages 5-7)	0.06	0.02
Reading and writing	During pre-k elligibility (ages 3-4)	0.2	0.03
Reading and writing	After pre-k elligibility (ages 5-7)	0.04	0.03
Math	During pre-k elligibility (ages 3-4)	0.2	0.04
Math	After pre-k elligibility (ages 5-7)	0.1	0.03
Other activities	During pre-k elligibility (ages 3-4)	0.07	0.03
Other activities	After pre-k elligibility (ages 5-7)	0.07	0.03
Rules and routines	During pre-k elligibility (ages 3-4)	0.13	0.03
Rules and routines	After pre-k elligibility (ages 5-7)	0.1	0.03
Tracking learning	During pre-k elligibility (ages 3-4)	0.22	0.04
Paternal involvement	During pre-k elligibility (ages 3-4)	0	0.07
Paternal involvement	After pre-k elligibility (ages 5-7)	0.05	0.07
Parent-school involvement	After pre-k elligibility (ages 5-7)	0.01	0.03

Notes: This table provides the underlying point estimates and standard errors used to construct figure 6 in the main text. All source estimates and standard errors for this table are taken from Gelber and Isen (2013). Treatment effects were originally reported in effects sizes in Gelber and Isen (2013) so no additional adjustment was necessary.

Table A.7: Additional details for figure 7

outcome_var	highschool_college	effect_size	standard_error
Graduate	Highschool	0.06	0.03
Graduate on-time	Highschool	0.054	0.03
Graduate	College	0.052	0.034
Attend	College	0.054	0.029
Attend 2-year	College	0.03	0.028
Attend 4-year	College	0.059	0.03
Semesters	College	0.569	0.322

Notes: This table provides the underlying point estimates and standard errors used to construct figure 7 in the main text. In all cases, effect sizes in this table are constructed by dividing source estimates by the associated outcome mean. Source estimates, standard errors, and control group outcome means used to create effect sizes for college-related outcomes are taken from the 2SLS estimates in table III of Gray-Lobe et al. (2023). Source estimates, standard errors, and control group outcome means used to create effect sizes for highschool outcomes are taken from the 2SLS estimates in table IV of Gray-Lobe et al. (2023).

Table A.8: Additional details for figure 8

intervention	outcome_category	effect_size	statistically_significant
Bierman et al (2008)	Behavior	-0.21	yes
Bierman et al (2008)	Emotion	0.21	no
Bierman et al (2008)	Vocabulary / Language	0.15	yes
Bierman et al (2020)	Behavior	-0.2	yes
Bierman et al (2020)	Emotion	-0.25	yes
Clements et al (2011)	Math	0.72	n/a
Farver et al (2009)	Vocabulary / Language	0.4	yes
Klein et al (2008)	Math	0.571	yes
Lonigan et al (2011)	Vocabulary / Language	0.27	yes
Upshur et al (2019)	Behavior	0.04	n/a
Upshur et al (2019)	Emotion	0.12	n/a

Notes: This table provides the underlying point estimates and standard errors used to construct figure 7 in the main text. The intervention column cites the paper where the source estimate can be found. Statistical significance as reported here is from the authors preferred method of inference; in two cases, the studies do not discuss inference. For Klein et al. (2008), treatment effects in effect size units are computed manually from source treatment/control means and control standard deviations contained in table 3 of that paper. All other studies contained in this table report effect sizes directly. Source estimates for Bierman et al. (2008) are drawn from table 3 (vocabulary, emotion identification, and CST aggressive response) of that paper; source estimates for Bierman et al. (2021) are drawn from table 2 of that paper; source estimates for Clements et al. (2011) are drawn from in text discussion on page 145; source estimates for Farver et al. (2009) are drawn from table 3 (receptive vocabulary) of that paper; source estimates for Lonigan et al. (2011) are drawn from table 2 (expressive language) of that paper; and source estimates for Upshur et al. (2019) are drawn from table 4 (Spring CST-Pro and Spring EMT) of that paper.

Table A.9: Additional details for figure 9

intervention	outcome_variable	professional_development_type	effect_size	statistically_significant
Raver et al (2009)	Disruption score	Behavioral PD	-0.854	yes
Raver et al (2009)	Disconnection score	Behavioral PD	-0.333	no
Powell et al (2010)	Receptive language	Literacy PD	0.03	no
Powell et al (2010)	Alphabet / letter knowledge	Literacy PD	0.32	yes
Wasik et al (2006)	Receptive language	Literacy PD	0.73	yes
Wasik et al (2006)	Alphabet / letter knowledge	Literacy PD	-0.33	yes

Notes: This table provides the underlying point estimates and standard errors used to construct figure 9 in the main text. The intervention column cites the paper where the source estimate can be found. Statistical significance as reported here is from the authors preferred method of inference. For Raver et al. (2009), effect sizes are computed manually by dividing the point estimates from table 3 (PIPPS column) by the corresponding outcome standard deviations in table 1. All other studies contained in this table report effect sizes directly. Source estimates for Powell et al. (2010) are drawn from table 5 of that paper; and source estimates for Wasik et al. (2006) are drawn from table 1 of that paper.