NBER WORKING PAPER SERIES

EARLY CHILDHOOD EDUCATION IN THE UNITED STATES: WHAT, WHEN, WHERE, WHO, HOW, AND WHY

Elizabeth U. Cascio

Working Paper 28722
http://www.nber.org/papers/w28722

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2021

I thank Charlotte Driscoll for excellent research assistance. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Early Childhood Education in the United States: What, When, Where, Who, How, and Why
Elizabeth U. Cascio
NBER Working Paper No. 28722
April 2021
JEL No. H75,I24,I28,J24,N32

ABSTRACT

This chapter concerns the state of the literature on early childhood education (ECE) – formal programs offering group instruction for children younger than the standard eligibility age for public education. I describe how ECE programs can be convincingly evaluated and why they may or may not work to narrow gaps in well-being across the lifecycle. The methods, the findings from their application, and their proper interpretation rest critically on who participates and when and where that participation is situated. Because there is a great deal of variation in the answers to these questions even within the United States, I focus on the U.S. experience. Over the past two decades, we have made considerable progress understanding the impacts of large-scale ECE participation in the U.S., as the literature has moved away from – despite being still strongly influenced by – small-scale model interventions. And yet, there is still much to be learned about the long-term effects of participating in ECE programs operating at scale, the mechanisms linking large-scale ECE interventions to later-life well-being, and the effects of ECE quality conditional on ECE participation.

Elizabeth U. Cascio
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
elizabeth.u.cascio@dartmouth.edu

I. Introduction and Overview

Socioeconomic gaps in human capital emerge before children start formal schooling. So too do socioeconomic gaps in access to preschool education. This is true not only within countries but also across: Children in lower-income countries are less likely to attend preschool the year before formal schooling begins. These realities have prompted a great deal of policy interest in expanding targeted formal learning opportunities to disadvantaged preschool-aged children, if not for making free preschool available to all children regardless of background, in effect lowering the grade at which public education begins.

This chapter describes how early childhood education (ECE) programs can be convincingly evaluated and why they may or may not work to narrow gaps in well-being across the lifecycle. The methods, the findings from their application, and their proper interpretation rest critically on who participates and when and where that participation is situated. Because there is a great deal of variation in the answers to these questions even within the United States, this chapter will focus on the U.S. experience. However, I will provide a selective review of the literature from elsewhere in the world, especially where it has been more advanced or innovative or offers insights that the extant variation or data in the U.S. have not yet been able to provide.

What is ECE, for the purposes of this chapter? I define ECE as formal programs offering group instruction for children younger than the standard eligibility age for public education.[1] This is a broad definition in some respects. For example, I include programs that are not funded or administered by public school systems, and indeed programs that are not publicly funded at all. However, it is narrow in another respect. By focusing on programs offering group *instruction*, I will be more selective in my review of evidence on programs offering group *care*,

[1] The standard age of eligibility can change over time. In the U.S., for example, it has gone from age 6 (before widespread availability of public-school kindergarten) to age 5, and now to age 4 in some states and localities.

i.e., childcare programs. This is not to say that such programs are unimportant: As mothers of young children have entered the labor force, provision of childcare during work hours has often been the explicit goal, rather than child development, and in practice, ECE programs always provide both care and instruction.[2]

Over the past two decades, we have learned a great deal about the impacts of ECE on the development of young children, their academic achievement and socio-emotional well-being as they progress through school, and their social and economic stature as adults. The credibility of this literature has benefited from small-scale randomized controlled trials, field experiments, and quasi-experiments arising from program implementation, as well as idiosyncratic features of programs already implemented at scale. In Section II, I describe the current ECE landscape and the evolution of ECE access in the U.S. Section III outlines the fundamental identification problem, the role of randomized trials and social experiments in addressing it, as well as approaches to translating policy shocks and other sources of access and attendance variation into research designs for ECE evaluation. Section IV reviews findings from the U.S. literature.

In Section V, I then offer and apply a simple conceptual framework to characterize the findings from the U.S. literature. The central insight is that impacts of any given ECE program on child development should be directly proportional to the quality of the learning environment it offers relative to what the average participant would have experienced in the absence of the program. This is an obvious deduction from the Rubin Causal Model (Holland, 1986): Treatment effects depend fundamentally on "potential outcomes" – the counterfactual – not just the treatment itself. But perhaps more in the ECE literature than in other settings, we commonly find

[2] My definition is narrow in another respect, in not incorporating either early childhood interventions within the home or other shocks to early childhood environments. For detailed reviews of these literatures, see Almond and Currie (2011) and Almond, Currie, and Duque (2018). For a recent alternative review of the ECE literature, see Elango, et al. (2016). In addition, Cascio (2015) provides a synopsis of universal ECE across the world.

that treatment effects rest as much on the counterfactual as they do on the program itself. Program quality is thus not revealed by the magnitude of a treatment effect, even though it is a tempting conclusion. I close this section with a brief review of the emergent literature on the impacts of ECE quality conditional on ECE attendance, or the "ECE production function." As an increasing number of children participate in ECE programs, the attendance margin becomes less important to understand than the quality one.

A weakness of the conceptual framework in Section V is that it does not offer different predictions about the impacts of ECE over the short- and long-term, despite the common finding of (sometimes rapid) "fadeout" in achievement effects of ECE as children age, but positive effects on non-test outcomes in late adolescence and adulthood. In the first half of Section VI, I, therefore, discuss alternative hypotheses for this pattern of findings and review of the literature attempting to distinguish among them. The second half of the section then describes our knowledge of alternative mechanisms through which ECE might affect a child's well-being, such as by changing maternal labor supply and parenting practices. Section VII closes with a rough guide to where research on ECE might fruitfully go in the future.

II. Who, When, and Where: The ECE Landscape in the U.S.

Publicly funded ECE programs in the U.S. have historically focused more on education and child development than childcare.[3] This section describes how the U.S. ECE landscape has evolved over the past century, emphasizing the public programs, social forces, and empirical regularities that have affected study design and interpretation and highlighting some data sources that have figured prominently in the literature.

[3] There are two key exceptions from emergency situations in U.S. history. During the Great Depression, the Works Progress Administration (WPA) funded nursery schools, and the 1940 Lanham Act funded childcare for working mothers during World War II (for the latter, see Herbst (2017)).

A. Trends in ECE Enrollment

Figure 1 shows trends in overall school enrollment in the U.S. (public and private schools combined) by age since 1920, based on public-use microdata from the Census (1920 to 2000) and American Community Survey (ACS) (2006 to 2018).[4] The school enrollment of 3- through 5-year-olds has grown a great deal over the past century. Enrollment rates of 5-year-olds have risen particularly strikingly – from less than 20% in 1940 to over 85% in recent years – with changes concentrated in the 1940s, 1960s, and 1970s. Among 3- and 4-year-olds, changes in enrollment have been smaller overall; by 2018, for example, the enrollment rate of 4-year-old children (3-year-old children) had only reached 61% (34%). However, these changes have also been more recent, concentrated in the 1970s and 1990s.

Growth in state funding of kindergartens (for 5-year-olds) and pre-kindergartens (mainly for 4-year-olds but also serving some 3-year-olds) contributed to these enrollment changes. Figure 2 shows ratios of kindergarten and pre-kindergarten (pre-K) to first-grade enrollment in public schools from 1939 through 2018 (right axis), based on data long published by the federal government, most recently through the Common Core of Data (CCD).[5] The kindergarten-to-first grade enrollment ratio exhibited a roughly constant growth rate between 1939 and 1980, rising from 0.2 to 0.9 (right axis). While kindergarten funding dates are not universally known before 1960, changes in the ratio after 1960 track state subsidization of school district provision of kindergarten (left axis).[6] The same is true for public school pre-K: The pre-K-to-first enrollment ratio began to increase faster after state funding of pre-K began in earnest in the mid-1980s.

[4] I use sampling weights to calculate enrollment rates and limit attention to individuals residing in the 50 states or Washington, D.C. who do not have allocated values for age or school enrollment.
[5] Importantly, the enrollment *ratio* is not the same as an enrollment *rate*, since cohort sizes may vary over time, and children may not progress through school normally (i.e., they may repeat or skip grades). As noted, school-based pre-K programs may also serve both 3- and 4-year-olds (Friedman-Krauss, et al., 2020).
[6] The last state to subsidize public school kindergartens was Mississippi, in 1983. See Cascio (2009a, 2009b).

The growth in overall enrollment shown in Figure 1 does not owe exclusively to expansions in public ECE; demand for private early education, fueled in part by rising rates of maternal labor force participation and thus the demand for childcare, also played a role. Figure 3 gives school enrollment rates both overall and in public schools only, calculated from the School Enrollment Supplements of the October Current Population Survey (CPS) from 1968 to 2018. A benefit of calculating enrollment rates with the October CPS is that a child's age in October – rather than on April 1, in the Census – is more aligned with the grade in which s/he should be enrolled, since most states and local governments have typically required entering kindergartners to be 5 years old sometime in the first four months of the school year.[7] As a result, overall enrollment rates at a given age are higher in the October CPS than the Census or ACS.[8] Figure 3 also shows that, though the expansion of state-funded pre-K has diminished the contribution of private schools to overall enrollment over time, private schools continue to play an especially important role in school enrollment under age 5. Rising private enrollment was also the main driver of enrollment growth among 3- and 4-year-olds in the 1970s.[9]

Figure 3 also shows a non-trivial public enrollment rate for 4-year-olds in 1968, before any states began to subsidize pre-K. The federal Head Start program may be responsible. Established in 1965 as part of President Lyndon Johnson's War on Poverty, Head Start targets the healthy development of preschool-aged children living at or below 130% of the federal

[7] These dates have changed over time (e.g., to September 1 from dates later in the fall) so that fewer 5-year-olds today (as of April 1) would have been eligible to start kindergarten the prior fall. See Deming and Dynarski (2008).
[8] For example, nearly half of children age 5 in October will have turned age 6 by April 1. This is one reason Figure 1 shows rising enrollment rates at age six in the Census over the past century: Many of these children are actually kindergarten age. Nevertheless, Census timing is only consistent (at April 1) across 1930 to 2000, and even across that period, there are differences in the "school referent": The 1930 Census asked about school enrollment as of September 1, the 1940 Census as of March 1, and the 1950 to 2000 Census as of February 1. In addition, the ACS is fielded throughout the year, asking about enrollment within the past three months.
[9] Another benefit of the October CPS over the Census/ACS is that it distinguishes between full-day and half-day kindergarten and preschool programs. Full-day kindergarten enrollment (regardless of age) outstripped half-day enrollment in 1995, and today approximately three-quarters of kindergartners attend full-day (Gibbs, 2017).

poverty line (FPL), by way of a federal matching grant to local governments and non-profits. Today, it largely serves 3- and 4-year-olds, but historically it also enrolled 5-year-olds where kindergartens were not offered (Zigler and Muenchow, 1992). To give a sense of the timing of changes in the scale of Head Start, Figure 2 shows the ratio of Head Start enrollment (across all age groups) to public school first-grade enrollment.[10] The high ratio in Head Start's earliest years reflects the fact that summer programs accounted for a large share of enrollment at its inception. Summer programs were phased out quickly later in the 1960s, though, and the ratio stabilized at about 0.11 in the 1970s. It rose slightly in the 1980s and more in the 1990s, due to Head Start funding expansions that decade.

B. Heterogeneity in ECE Enrollment

The ECE enrollment changes described above have been unevenly felt across various segments of American society. In part, this is by design: Head Start targets low-income children, as do several state-funded pre-K programs. However, geographic variation in when states began funding kindergartens in particular, as well other forces increasing demand for private programs, such as rising maternal employment, have mattered as well.

Table 1 presents gaps in overall school enrollment rates at ages 3 through 5 across region, race, and two measures of family background – whether the maximum education of a child's parents is at or below the median (amongst all parents with young children in a given year) and whether a child's mother is currently working.[11] Despite the misalignment between age and grade of enrollment, I use data from Census and ACS to provide the longest possible perspective,

[10] Because it incorporates multiple age groups in Head Start, the Head Start-to-1st grade enrollment ratio overstates enrollment rates at any given age.

[11] Sample splits by parental characteristics are based on parents residing in the same household as their children at the time of the survey; parental characteristics are also measured at the time of the survey. Characteristics of non-resident parents or at a fixed child age are not observable in the Census and ACS. They are observable in other longitudinal data sets described in Sections III and IV.

spanning 1940 (the first year with data on educational attainment) to 2018. Table 2 is structured like Table 1 but focuses on public school enrollment, which the Census reports starting in 1960.

The South lagged the rest of the country in enrollment of 5-year-olds both overall and in public schools (Panel A of each table). Though the South-non-South gap in overall enrollment at age 5 was already substantial in 1940, it grew between 1940 and 1960 before narrowing thereafter, as states across the South began subsidizing school district provision of kindergarten. Thus, big regional differences underlie the aggregate trends in age 5 enrollment shown in Figure 1 and in public school kindergarten enrollment and funding shown in Figure 2. In terms of overall enrollment, this regional variation in timing had little impact on race gaps (Table 1 Panel B), but did translate into widening, then narrowing, gaps by parental education (Table 1 Panel C), reflecting the relatively low levels of parental education in the South and that southern children regardless of race were unlikely to attend kindergarten before it was funded. Enrollment of 5-year-olds overall is still higher among children whose parents have more education, though these children are relatively less likely to be in public school. Enrollment of 5-year-olds is also higher among children whose mothers work, owing essentially entirely to higher private school enrollment (Panel D).

Despite the fact that public enrollment at younger ages also favors disadvantaged children, overall enrollment gaps by parental education at ages 3 and 4 are much larger. They have changed little since 1980 (Panel C of Tables 1 and 2). In 2018, 4-year-olds whose parents' maximum educational attainment was at or below the median were 18 percentage points less likely to be enrolled in school; for 3-year-olds, this gap was even larger, at 22 percentage points. Part of the relatively high demand for private schooling among more educated families at these young ages may stem from a need for childcare, as overall enrollment rates are also higher

among 3- and 4-year-olds whose mothers work (Table 1 Panel D), and maternal employment is positively correlated with maternal educational attainment.[12] But it may also reflect the greater resources available in these families for human capital investment.

Consistent with the observations regarding private schooling at ages 3 and 4 in Figure 3, the enrollment gap between young children whose mothers are employed versus not expanded in the 1970s. Figure 4 shows trends in maternal employment by child age from 1930 to 2018, based on public-use microdata from the Census and ACS. As expected, maternal employment is more common among the mothers of older children. However, much more striking are the common time trends across age groups – a pattern much different than that in Figure 1. The employment of women with young children grew steadily between 1940 and 1990, with growth accelerating in the 1970s. But since 1990, maternal employment in the U.S. has grown by much less, reaching between 62 and 67 percent for mothers of 3- to 7-year-olds in 2018.

C. Comparison of the U.S. to Other Countries

Figure 5 compares U.S. school enrollment rates at age 4 to those in other countries in 2017, the latest year when data from a large number of countries are available from UNESCO. The U.S. ranks at the bottom of OECD countries (Panel A), with a pre-primary school enrollment rate only above that of Turkey. The establishment and continued operation of universal childcare and preschool programs among OECD countries at or near the top of this ranking – Germany, Norway, Denmark, among others – have greatly informed our understanding of early childhood education and care. This also has been the case for several countries in the Americas (Panel B). I offer a selective review of this evidence, as well as note some of the

[12] The correlation is however not strong; for example, the correlation between indicators for a mother of a three- or four-year-old working and having a high school (college) degree was 0.14 (0.08) in 1980 and 0.17 (0.17) in 2018 (author's calculations from Decennial Census and ACS public-use microdata).

research innovations happening in the context of developing countries, in the remainder of this chapter.

III. How: Research Designs to Evaluate ECE Programs

The fundamental challenge to ECE evaluation is that children are selected into program participation, possibly based on unobservable determinants of outcomes. As a result, a difference in the average outcomes of attendees and non-attendees in observational data – even if regression-adjusted for differences in their observed characteristics – will reflect the effects of unobserved characteristics, in addition to the effects of program attendance. The sign of the resulting bias depends on the setting or program. For example, Head Start attendees are, on average, quite disadvantaged due to program eligibility rules. In this case, the likely bias is negative: Those children might still perform worse on tests or in life than children who do not attend, even if Head Start makes them better off. On the other hand, less disadvantaged children are more likely to attend a private preschool. Here, the likely bias is positive: These children are likely to be better off later in life, even in the absence of preschool attendance.

This section reviews common empirical approaches for addressing this central identification problem in ECE evaluation. I present the approaches in order of the strength of their theoretical *internal validity*, or credibility of the causal inferences they generate and offer some key examples of where they have been used in the literature. However, all of the approaches – even experiments – have limitations. In particular, there can be (but is not necessarily) a trade-off between internal validity and *external validity*, or generalizability.

A. Randomized Controlled Trials

The so-called "gold standard" of program evaluation is the randomized controlled trial (RCT). In an RCT, the treatment – here, preschool or ECE attendance – is assigned randomly, or

in effect via successive flips of a fair coin or lottery. With the random assignment of ECE attendance, attendees will on average have the same characteristics as non-attendees, i.e., treatment and control groups will be "balanced" on other characteristics, aside from attendance, that bear on outcomes of interest like test scores and high school completion. If an RCT is well-implemented, a simple difference in average outcomes between the treatment group and the control group thus identifies the average effect of ECE attendance – the average treatment effect (ATE).

Two ECE RCTs carried out in the U.S. in the 1960s and 1970s generated core early knowledge on the impacts of ECE attendance. In the mid-1960s, the High-Scope Perry Preschool Project (PPP) randomly assigned about half of around 120 low-income, low-achieving 3 and 4-year-olds children in Ypsilanti, Michigan into a morning preschool program during the school year, augmented with home visitation; the remaining children placed in the control group experienced business as usual. Researchers have attempted to follow all initial participants over time, so the PPP has been widely studied (e.g., Schweinhart et al., 2010; Heckman et al., 2010; Heckman, Pinto, and Savelyev, 2013). Similarly, the 1970s saw initial randomization of participation in the Carolina Abecedarian (ABC) program, which offered full-time, full-year formal care from early infancy to school age to a random subset of 110 recruited participants in the area of Chapel Hill, NC.[13] ABC participants have also been followed over time and have been the subject of a large literature (e.g., Campbell et al., 2014).

Why have more ECE RCTs not been carried out? The expense of RCTs may necessitate small numbers of participants. Small RCTs may not have balanced treatment and control groups at the outset, or they may be quick to experience selective attrition as participants age. Though

[13] The original ABC demonstration program also included a program for 5- to 8-year-olds. All discussion of ABC in this chapter concerns the intervention between ages 0 to 5.

there are empirical approaches for addressing these issues, they can rely on strong assumptions (e.g., selection on observables in the case of propensity score reweighting) and weaker assumptions may generate uninformative estimates (e.g., worst-case scenario bounds to address attrition may be wide). In the early years, it was also common to ignore multiple inference in PPP and ABC evaluation. Accounting for multiple inference makes the estimates less precise and less conclusive (Anderson, 2008), but conclusions can still be drawn (Elango et al., 2016).

In addition, the (potential) strong internal validity of RCTs may be compromised by weak external validity – concerns over the ability to generalize from the small and highly localized populations on which they are implemented. Compounding this concern is the fact that RCTs may be easier to carry out on pilots of new programs than on well-established programs. While we have learned much from these RCTs, as will be discussed later in the chapter, it has been important from a policy perspective to consider programs operating at scale.[14]

B. Social Experiments

There is a middle ground between an RCT and a quasi-experiment in the ECE literature: social experiments, or settings where the opportunity to attend an ECE program – rather than attendance itself – is randomly assigned. The implication of randomization is analogous to the RCT case: The treatment group (those offered a slot) should be on average identical to the control group (those not offered a slot). However, a social experiment is easier to carry out on a large and diverse population and a program operating at scale.

[14] Unlike the model demonstrations of PPP and ABC in the U.S., some RCTs in low- and middle-income countries have involved larger and more diverse samples. Even though these programs have themselves been novel for the setting, they have been implemented as outgrowths of existing social service infrastructure. A case in point is a program in Colombia that provided random subsamples of approximately 1400 toddlers across 96 municipalities home-based psychosocial stimulation, micronutrient supplementation, or both (e.g., Attanasio, et al., 2014). Other parenting and micronutrient interventions have taken place throughout the developing world, e.g., in Jamaica (Grantham-McGregor et al., 1991, 2007; Gertler, et al., 2014). Strictly speaking, these programs do not meet my definition of ECE, since they do not involve formal group-based care or education outside of the home.

Another key difference between a social experiment and an RCT, however, is the potential for "non-compliance": Not all of the treatment group will accept the offer, and some of the control group will participate in the ECE program (even if not at the same site) in the absence of an offer. Thus, unlike in the pure RCT case, the difference in average outcomes between the treatment and control group does not identify the ATE; rather, it identifies the intent-to-treat (ITT) effect. The ITT, when scaled up by the difference in attendance rates across the treatment and control groups, then identifies the effect of the treatment on the treated (TOT), or the local average treatment effect (LATE).[15] The TOT/LATE can be estimated via two-stage least squares (2SLS) regression on the study microdata, instrumenting for an attendance indicator with a dummy for the randomized offer of a slot, controlling for site fixed effects in a multi-site setting with within-site randomization.

The TOT/LATE captures the impact of ECE attendance for the subset of applicants incentivized to participate in the ECE program due to offer assignment. If the true model is one with heterogeneous effects of participation across individuals – some people benefitting more than others – the TOT/LATE and the ATE will not be the same. For one, the estimation sample in a social experiment consists of the subset of the population that applies to participate in a particular program, and applicants might have different expected gains from participation. Researchers typically argue that the TOT/LATE in the context of a social experiment is still useful and interesting since it captures the impact of program participation for individuals both motivated to apply for a program and incentivized to attend via intervention.

[15] This is only true under the assumption of monotonicity. Monotonicity requires that there be no "defiers" – individuals who take up the program when they are not offered a space, but do not take it up when they are offered one. Importantly, observing that some people take up the program when they are not offered a space is not necessarily a violation of monotonicity: these individuals could be "always takers," or individuals who always take up the program regardless of treatment status. Nor is it necessarily a violation of monotonicity to observe some treated individuals not taking up the program; they could be "never takers." See Angrist, Imbens, and Rubin (1996).

Social experiments have recently gained traction in ECE evaluation. Perhaps most famously, the Head Start Impact Study (HSIS) was a congressionally mandated randomized evaluation of Head Start. Offers to participate were randomized across nearly 5,000 3- and 4-year-old applicants to over 300 over-subscribed Head Start sites across the U.S. in fall 2002, and participants have been followed through third grade. The resulting data have been subject to internal evaluation by the Department of Health and Human Services (HHS) (e.g., Puma et al., 2010) and re-evaluation by academics (e.g., Kline and Walters, 2016; Feller et al., 2016). In a similar spirit, Dean and Jayachandran (2019) randomized scholarships to attend private kindergarten across around 800 children in 71 villages in Karnataka, India, to estimate the short-term impacts of private kindergarten attendance. Just as in the HSIS case, compliance was imperfect, necessitating a 2SLS approach to identify the TOT/LATE.

In low- and middle-income country contexts, it has also been possible to randomize the roll-out of publicly funded preschool. From a methodological perspective, the randomized establishment of a program locally can be thought of as much the same as randomization of an offer. Bouguen et al. (2018) evaluate a preschool construction program in Cambodia, which randomly assigned early implementation of preschool as part of the local primary school in 26 of 45 villages. Likewise, Martinez, Naudeau, and Pereira (2017) exploit the random establishment of community preschools across 30 of 76 villages in rural Mozambique. In both of these studies, researchers have access to survey panel data on at least 1500 to 2000 children, so they were carried out on a similar scale as the other social experiments described above.

With the potential to deliver both internally and externally valid estimates, social experiments could be thought of as the "sweet spot" for ECE evaluation. When they reach sufficient fidelity of implementation, they can generate deeper insights. For example, researchers

have used the data from the HSIS to explore the correlates of cross-site heterogeneity in program impacts (Walters, 2015), to estimate treatment effects under alternative counterfactuals (Kline and Walters, 2016; Feller et al., 2016), and to estimate treatment effects on test performance across the distribution (Bitler, Hoynes, and Domina, 2014).

C. Quasi-experiments

Unlike RCTs and social experiments, quasi-experiments seek to exploit "as good as random" variation in observational data. There are three key quasi-experimental approaches in the ECE literature: (1) regression discontinuity designs exploiting program rules, often regarding eligibility; (2) difference-in-differences designs typically exploiting program implementation; and (3) family fixed effects designs exploiting extant variation in ECE attendance across siblings. Like social experiments, (1) and (2) have to be combined with 2SLS to recover the TOT/LATE. In principle, the ATE can be recovered from (3).

1. Eligibility Criteria: Regression Discontinuity

Aside from social experiments, another way to evaluate ECE programs operating at scale is to take advantage of idiosyncratic features of eligibility criteria. For example, as earlier noted, most U.S. ECE programs have strict birthday cutoffs for eligibility, such as the requirement in some states (or localities) that children entering pre-kindergarten be 4 years old by August 31. Targeted ECE programs may also feature family income cutoffs, such as Head Start's requirement that 90% of its enrollment be comprised of children from families living at or below 130% FPL. The consequence of the former is a large difference in current-year preschool participation rates of children who are basically the same age; the consequence of the latter is a large difference in participation rates across children from families of similar means.

In both of these examples, there is no explicit, researcher-controlled random assignment of an offer to participate. However, program rules can, in effect, generate random variation in offers among children in certain parts of the age and income distributions. In the former case, for example, offers are "as good as random" across children turning age 4 in (late) August (the treatment group) and (early) September (the control group) of a given school year if children who are basically the same age are on average the same in other respects that might affect their test performance later in the year. In the latter case, offers of attendance are as good as random across children whose families have incomes close to 130% FPL if these children are the same as each other, on average, in terms of their latent talent.

The parallel to the social experiment case would be to identify the TOT/LATE by scaling the local (around the age or income threshold) difference in mean outcomes by the local difference in participation rates, using 2SLS with an eligibility indicator as an instrument for attendance. However, in most data, there are too few observations right around the threshold for such an estimate to be informative. One answer to this challenge would be to widen the age or income span of observations included in the sample. Instead of restricting attention to children with late August or early September birthdays, for instance, one could consider children born between June and November. However, this would compromise (as good as) random assignment: in the cross-section, children turning age 4 between June and August are on average significantly older than those turning age 4 between September and November, and small differences in age matter considerably for developmental outcomes among young children (see, for example, Elder and Lubotsky, 2009; Cascio, 2020). These children may also differ in terms of other background characteristics that matter for test scores.

The regression discontinuity design (RDD) provides a way of using more data to predict outcomes in the neighborhood of the threshold value of the running variable (age or family income) that determines treatment (program eligibility). In the RDD, the relationship between the running variable and outcomes is modeled as smooth in the absence of treatment. That is, age can have a positive effect on test scores; some polynomial function must just capture the relationship between age and test scores. If there is a treatment effect, this relationship should be discontinuous at the threshold. In a so-called "fuzzy" RDD, where attendance does not rise one-for-one with eligibility, the magnitude of this discontinuity will in principle identify the ITT. The TOT/LATE can then be identified by instrumenting for attendance with an eligibility indicator, conditional on the smooth function in the running variable, using 2SLS.

The RDD has become a popular tool in ECE evaluation. For example, the age-eligibility RDD has been successfully applied to estimate the impacts of pre-K and kindergarten eligibility and attendance on maternal labor supply in the U.S. (Fitzpatrick, 2010, 2012), Argentina (Berlinski, Galiani, and McEwan, 2011), and Germany (Bauernschuster and Schlotter, 2015). It has also been applied to estimate impacts on child test scores in studies of the pre-K programs in Tulsa, Oklahoma (Gormley and Gayer, 2005), Boston (Weiland and Yoshikawa, 2013), and Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia (Wong et al., 2008). However, limitations in the administrative data on which this second set of evaluations has relied, such as censoring of children who do not enroll in a public school at the start of the school year, make it impossible to interpret their estimates as ITT effects in the purest sense. For the same reason, their 2SLS estimates using eligibility as an instrument for attendance also do not necessarily identify the true TOT/LATE.[16]

[16] Lipsey et al. (2015) describe these and other limitations of the age-eligibility RDD as typically implemented in practice with school administrative data.

As alluded, the ECE literature also features the use of RDD to exploit other eligibility criteria. For example, Carneiro and Ginja (2014) take advantage of Head Start's income-eligibility rules to identify the effects of the program on later-life health and behavioral outcomes. Also in the context of Head Start, Ludwig and Miller (2007) use the fact that, at the program's start, the Office of Economic Opportunity (OEO) in the Department of Health, Education, and Welfare (HEW) targeted the poorest 300 counties for additional assistance in preparing their applications. Even after the actual assistance had ended, counties that had been just barely eligible for it had greater Head Start penetration and higher Head Start attendance rates than counties just barely missing eligibility. Ludwig and Miller (2007) exploit this discontinuity in treatment with county poverty to estimate the impact of Head Start on child mortality and educational attainment.

Like small-scale RCTs, studies using the RDD tend to have relatively strong internal validity, but potentially weak external validity.[17] Unlike in the RCT case, however, weak external validity rarely stems from a failure to capture the impacts of a program operating at scale, but rather from the very narrow subset of the population that forms the basis of identification. In a setting with heterogeneity in impacts, treatment effects identified from attendance variation near the eligibility threshold might not be broadly representative.

2. Program Implementation: Difference-in-Differences

As noted, it has been possible to randomize the roll-out of ECE programs across areas in some developing countries. However, this is not common. Programs are more often established in different areas at different points in time, not by design but rather due to cross-area variation either in the timing of responses to a mandate or of subsidization from some higher level of

[17] Importantly, strong internal validity rests on an inability to manipulate the running variable to affect eligibility. This must be a consideration in any future use of RDD.

government or in preferences for and constraints on establishment of locally funded programs. Still, this variation across areas and over time, such as that described in Section II in the context of kindergarten and pre-K, presents an opportunity for identification. Areas without a change in ECE supply from a specific year to the next can potentially serve as a comparison group for those that do, with changes over time in outcomes in the comparison group providing an estimate of what would have happened in treated areas in the absence of the expansion in supply. The comparison group thus takes the place of the control group in an RCT or social experiment.

This is a difference-in-differences (DD) approach: the first difference is the change over time (or across cohorts) in treated areas, whereas the second difference is the change over time (or across cohorts) in comparison areas. In settings where only one area implements a program, as was the case with Georgia's early implementation of universal pre-K (Fitzpatrick, 2008) or Quebec's universal childcare program (e.g., Baker, Gruber, and Milligan, 2008, 2019), the baseline DD model can be straightforward, including dummies for the treatment area, time or cohort, and their interaction, with the coefficient on the interaction being the DD estimate. In cases where multiple geographies implement the same basic program but at different points in time, such as with kindergartens for 5 year-olds in the U.S. (Cascio, 2009a, 2009b) or childcare for 3- to 6-year-olds in Norway (Havnes and Mogstad, 2011a, 2011b, 2015), the baseline DD model includes a post-implementation indicator, which turns on at different times (or different cohorts) in different areas, along with area and time (or cohort) fixed effects. The DD estimate is the coefficient on the post-implementation indicator, which is a weighted average of all possible two-by-two DD estimates that can be constructed from the data (Goodman-Bacon, 2018).

DD has been the workhouse quasi-experimental approach of the ECE evaluation literature not just for U.S. programs, but for programs worldwide. The approach has been used to

estimate the impacts of Head Start (Bailey, Sun, and Timpe, 2020; Barr and Gibbs, 2019; Thompson, 2018; Johnson and Jackson, 2019), kindergarten and pre-K in the U.S. (Cascio, 2009a, 2009b; Fitzpatrick, 2008; Cascio and Schanzenbach, 2013), preschool programs in Argentina (Berlinski and Galiani, 2007; Berlinski, Galiani, and Gertler, 2009) and England (Blanden et al., 2016), and childcare programs in Quebec (e.g., Baker, Gruber, and Milligan, 2008, 2019; Lefebvre and Merrigan, 2008; Haeck et al., 2015), Norway (e.g., Havnes and Mogstad, 2011a, 2011b, 2015), Germany (e.g., Bauernschuster and Schlotter, 2015), and Spain (Felfe, Nollenberger, and Rodriguez-Planas, 2014; Nollenberger and Rodriguez-Planas, 2015), among others. Some of these studies scale up reduced-form DD estimates for outcomes with first stage DD estimates for ECE attendance, either informally or formally through 2SLS.

One challenge with the DD design is that the timing of area implementation is typically not random. The industry standard approach to this concern has been to estimate "event-study" models, which replace the post-implementation indicator in the baseline model with a series of indicators for "event time," or year relative to the year before implementation. The ideal finding from event-study estimation is that the coefficients on the pre-implementation event time indicators are close to zero in magnitude and not statistically significant. Dynamic treatment effects can, however, complicate interpretation, when early adopters for which treatment effects are still phasing in are used as controls for later adopters (Goodman-Bacon, 2018). With the exception of a recent paper (Zerpa, 2020), the DD literature in ECE has not yet attempted to reconcile with this critique.

3. Sibling comparisons: Family fixed effects

The earliest quasi-experimental ECE evaluations compared siblings with different ECE experiences. Most famously, a series of papers by Janet Currie and Duncan Thomas exploit

variation in Head Start attendance and preschool participation conditional on mother fixed effects to estimate the impacts of Head Start (Currie and Thomas, 1995; Currie and Thomas, 1999; Garces, Thomas, and Currie, 2002). Subsequently, Deming (2009) used sibling comparisons to estimate the impacts of Head Start attendance on a host of outcomes from childhood through adolescence, and Pages et al. (2020) have extended outcomes through adulthood. Outside of Head Start, Berlinski, Galiani and Manacorda (2008) combine sibling comparisons with a rapid expansion of pre-primary education to estimate the impact of preschool attendance on school progression in Uruguay.

A nice feature of family fixed effects (FFE) models is that they can be readily estimated from surveys that collect data on all individuals (or all individuals meeting certain age criteria) in a sampled household. For example, the Panel Study of Income Dynamics (PSID), used in Garces, Thomas, and Currie (2002), attempts to survey all individuals within all households subsequently formed by members the original PSID households in 1968. The Children of the National Longitudinal Survey of Youth (CNLSY), used in Currie and Thomas (1995, 1999), Deming (2009), and Pages et al. (2020), includes all children born to women in the original (1979) NLSY (NLSY79). These are rich longitudinal data sets with many outcome variables in addition to reports of Head Start attendance.

Despite the ease of implementation, there are drawbacks to the FFE approach. First, differences in the ECE experiences of siblings are not necessarily random. While it is common to include rich person-level controls in FFE models to account for non-random selection into program participation, few studies show that there is balance on observables within families (e.g., Deming, 2009), or attempt to combine policy variation with siblings comparisons (e.g., Berlinski, Galiani, and Manacorda, 2008). Second, in some surveys, like the PSID, reports of

Head Start attendance are retrospective, raising the possibility of attenuation bias from misclassification. Garces, Thomas, and Currie (2002) present a variety of evidence to suggest that misclassification is limited, but this has not yet been validated through any merge of survey data to administrative records. Third, Miller, Shenhav, and Grosz (2019) show that siblings comparisons rely heavily on larger families for identification, thus making the estimates more representative of larger families than the population as a whole. When FFE estimates of Head Start attendance in the PSID and NLSY are reweighted to reflect the true distribution of family size in the population, ATE estimates shrink considerably.

IV. Overview of the U.S. ECE Evaluation Literature

This section provides an overview of the findings of the U.S. ECE evaluation literature. I organize the overview by the program, rather than by the outcome variable, which might be more standard for a chapter of this kind, since findings from a given program may be better understood holistically. Within program, I also organize the review by the cohorts of study. As described in Section II, the alternatives to program participation have changed dramatically over time due to changes in which other public programs might be available and other forces increasing private preschool demand, such as rising maternal employment. Informed by the literature, I formalize this idea with a simple framework in Section V.

A. Model Interventions in the U.S.

PPP and ABC have by now been subject to extensive study. Because these RCTs were carried out so long ago, it has been possible to follow study participants throughout childhood and well into adulthood. Findings from these RCTs provide the foundation of what we know about ECE and a benchmark to which to compare findings from the literature on larger-scale

ECE programs. I offer a limited and selective review of this literature here since it has been covered extensively elsewhere.

Elango et al. (2016) provide a recent synopsis of findings from PPP and ABC, addressing common criticisms of small-scale RCTs, such as multiple inference. They report that participants in both demonstration programs experienced substantial initial gains in IQ. The PPP treatment group scored on average 11.4 points – or 0.75 (population) standard deviations (s.d.) – higher on an IQ test at age 5 than their control counterparts. For ABC, the treatment effect on IQ at age 5 was smaller but still substantial, at 6.4 points, or 0.42 s.d. But these effects faded as children aged; by age 8, the impact on IQ in PPP was almost gone (1.25 points or 0.08 s.d.), though it remained substantial in ABC (4.5 points or 0.3 s.d.). Still, effects on achievement test scores at school age, measured as factors based on the California Achievement Test (CAT), the Peabody Individual Achievement Test (PIAT), and the Woodcock-Johnson Test of Achievement (WJAT), are substantial for both programs, at 0.4 s.d. (PPP) and 0.5 s.d. (ABC).

Despite fadeout in IQ effects, both PPP and ABC improved later-life outcomes. For females, PPP delivered increases in high school graduation (as of age 19) and reductions in the likelihood of having ever been arrested (by age 40) or on welfare (ages 18 to 27); for males, it reduced cumulative arrests as age 40 and increased employment (as of age 40).[18] By contrast, ABC increased educational attainment regardless of sex, as well as reduced cumulative arrests (by age 34) only for women; it also raised women's earnings. Similar to PPP, however, ABC increased male employment. It also reduced male obesity and hypertension (outcomes not measured in PPP). Heckman et al. (2010) offer a painstakingly careful and thorough cost-benefit

[18] A recent working paper (Heckman and Karapakula, 2019b) shows significant treatment-control differences at age 55 in PPP, applying new statistical methods to address "incomplete knowledge of and compromises in the randomization protocol used to form the control and treatment groups," as well as non-response and attrition.

analysis of PPP, concluding that its benefit-cost ratio ranges between 3.9 and 6.8 to one. A thorough cost-benefit analysis of ABC likewise yields a benefit-cost ratio of 3.2 to one (Elango et al., 2015).[19]

In addition to establishing a pattern of effects that has now become commonplace in the ECE evaluation literature – fadeout in cognitive effects, but persistence in later-life social and economic outcomes – both demonstration programs have also provided critical insights into *why* ECE works (see Section VI). The current age of participants has also made it possible to extend evaluation to the second-generation – effects on the children of participants. Heckman and Karapakula (2019a) show that the children of treated PPP participants were less likely to have been suspended in school or participate in crime and had higher levels of education and employment in their 20s. Effects were larger in the subsample of male participants.

B. Head Start

Unlike any other U.S. ECE program, Head Start has been evaluated with almost every research design outlined in Section III – a social experiment, RDD, DD, and FFE. Moreover, the same research design – FFE – has been applied to different cohorts of children, helping to shed light on the importance of a changing counterfactual to the estimates. Also, just like PPP, the program is now mature enough to provide insight into not only the long-run impacts of participation but also second-generation impacts.

As described in Section II, when Head Start was established in 1965, there were limited alternative formal learning opportunities for preschool-aged children, and maternal employment rates were lower than they would later become. As a result, the counterfactual for the typical

[19] This discussion has drawn heavily from findings reported in Tables 4 and 7 of Elango et al. (2016). I refer the interested reader to that paper, and to the studies on which their synopsis was based, specifically Heckman, et al. (2010), Heckman, Pinto, and Savelyev (2013), Campbell et al. (2014), and Elango et al. (2015).

Head Start participant would have been home or informal care. In addition, by design, the average participant was quite disadvantaged. To the extent that Head Start has an impact, we might therefore expect it to be greatest for the earliest exposed cohorts, assuming that program quality has not increased much over time. I thus review the Head Start literature roughly chronologically by the cohort affected.

1. Participants in the 1960s and 1970s

Several studies consider the short-term effects of Head Start for the earliest cohort of participants. Ludwig and Miller (2007) exploit the persistent differences in Head Start penetration and attendance between counties that were just barely eligible versus just barely eligible for special OEO grant-writing assistance in HEW at the program's inception, given their 1960 poverty rates. Their most robust finding concerns child health: Using Vital Statistics data from 1973 to 1983, they find evidence of reductions in child mortality between the ages of 5 and 9 for causes that could plausibly be related to the health services that Head Start provides. Using data from the National Collaborative Perinatal Project and variation in the intensity of Head Start funding across areas at the program's inception, Aizer and Cunha (2012) find larger gains in IQ scores for participants with higher human capital endowments.

Due to data limitations (e.g., lack of test score data for these cohorts), the rest of the knowledge base about the earliest Head Start participants is restricted to longer-term outcomes. Comparing siblings in the PSID who could have been of age to attend Head Start between 1966 and 1977, Garces, Thomas, and Currie (2002) conclude that the program boosted the high school completion, college-going, and earnings of white children and reduced the chances that Black children had been booked or charged with a crime. There are larger impacts for some outcomes among children whose mothers had no more than a high school degree. The FFE point estimates

are, however, substantial in magnitude. A re-evaluation of these data by Miller, Shenhav, and Grosz (2019) concludes that the ATE for college-going is small and not statistically significant. However, the confidence interval is wide enough to include effects consistent with a more recent wave of studies on these cohorts.

More recent studies have taken advantage of the timing and intensity of Head Start's roll-out across counties to estimate impacts on a host of longer-term outcomes across participants in both survey and administrative data. Using data from the NLSY79, in which respondents were born 1957 to 1964 (and thus eligible to attend Head Start through the late 1960s), Thompson (2018) finds that greater Head Start exposure – defined as average program funding in one's county of birth at ages 3 to 6 – was associated with greater own and household income, higher completed education (specifically college-going), and fewer health limitations in adulthood. Augmenting the NLSY79 with data from the CNLSY and using the same basic source of variation, Barr and Gibbs (2019) find that *having a parent* exposed to Head Start reduced a person's chances of becoming a teen parent or engaging in crime, and increased their chances of completing high school and going on to college. Their conclusions are maintained when they exploit the sharp difference across counties with similar poverty level levels in terms of OEO grant-writing assistance (Ludwig and Miller, 2007). While the two papers define disadvantage differently, they are consistent in finding larger effects for more disadvantaged populations.

Johnson and Jackson (2019) arrive at similar basic conclusions as Thompson (2018) but using the PSID, focusing on the program as it operated between 1965 and 1980. They instrument for Head Start funding at age 4 with the presence of a Head Start center, as well as estimate a specification that allows the impacts of Head Start to vary with later exposure to school finance reform (SFR) induced increases in school spending. This is a test of "dynamic complementarity"

– whether later investments augment the impacts of early investment (see Section VI). They find that Head Start funding is associated with higher levels of educational attainment, higher wages, lower poverty, and lower incarceration rates among poor children. The positive effects of Head Start are larger for those exposed to SFR.

A common feature of Thompson (2018), Barr and Gibbs (2019), and Johnson and Jackson (2019) is a reliance on relatively small-scale survey data. Two other recent studies apply county-by-year variation from the roll-out of Head Start to much larger scale administrative data. One chief drawback of administrative data is that outcomes have the potential to be more limited. Another is that nothing is known about actual program participation or parental characteristics beyond what can be inferred from county of birth. The main benefit of administrative data, however, is that they are many orders of magnitude larger than the PSID or CNLSY, which helps to reduce standard errors and thus uncertainty about effect sizes.

Bailey, Sun, and Timpe (2020) use non-public data on adult educational attainment, income, and public assistance from the 2000 long-form Decennial Census and the 2001-2013 ACS. Linking these data to the Social Security Administration (SSA)'s Numident file, they obtain two pieces of information that are helpful for estimating Head Start exposure – county of birth and exact birthdate, which they combine with school entry cutoff birthdates to assign children to school entry cohorts. Rather than just comparing exposed versus non-exposed cohorts in affected versus unaffected counties, their model allows treatment effects to be proportional to how long the local Head Start program has been in place, to reflect possible improvements in program quality and more years of potential attendance. Their findings suggest that Head Start exposure improved educational attainment and economic self-sufficiency. Moreover, where

outcomes overlap (high school graduation and college attendance) with those found in previous literature, their estimates are smaller. However, as anticipated, they are also more precise.[20]

Anders, Barr, and Smith (2020) examine the effects of Head Start on crime using North Carolina criminal convictions data that, like the linked data described above, include information on county of birth and birthdate. Their key outcome is the county-by-cohort conviction rate for serious (property and violent) crimes by age 35, and they focus on Head Start exposure at age 4. While they see no evidence of an effect when pooling across counties, they find that Head Start exposure reduced the serious criminal conviction rate in higher-poverty counties, where more children should have been eligible for Head Start. The magnitude of the estimates in higher-poverty counties is between that for crime found in the survey-based analyses of Garces, Thomas and Currie (2002) and Johnson and Jackson (2019) and is more precisely estimated.

Finally, de Haan and Leuven (2020) depart from the use of quasi-experimental variation, instead estimating lower bounds on the long-term effects of Head Start participation across the distributions of earnings and educational attainment. Their approach models counterfactual quantities (i.e., the impact of Head Start participation for a non-participant) by making weak stochastic dominance assumptions concerning outcome distributions conditional on parental education and Head Start participation. They present evidence that these assumptions are satisfied among non-treated cohorts in the NLSY79 before applying the method to treated cohorts. The approach reveals significant positive effects on education and wage income, but only at the bottom of the distribution. Effects are larger for Blacks and Hispanics. Like in Garces, Thomas, and Currie (2002), their Head Start treatment is attendance, not just exposure.

[20] Ludwig and Miller also use their RDD approach to estimate impacts on educational attainment in both survey data (the 1988 National Educational Longitudinal Survey) and in educational attainment at the county level in the Census, based on published and special tabulations that rely on more data than available in public-use microdata samples. These estimates are less conclusive than those for child mortality but suggest increases in attainment.

2. Participants in the 1980s and 1990s

There are fewer studies of Head Start's effects on later cohorts. This owes in part to the fact that the roll-out of Head Start, which provides the source of identification in many of the studies described above, could only happen once. However, the recent outpouring of work using the roll-out variation reflects an academic interest in program impacts on longer-term well-being; participants in the 1980s and 1990s are only now starting to be old enough to estimate impacts on educational attainment or earnings. Thus, studies of Head Start participants in the 1980s and 1990s have relied on research designs besides DD and until recently (Pages et al., 2020) been limited to short- and medium-term outcomes observed in survey data, particularly the CNLSY.

The favored research design for these cohorts is FFE. In the first study of this type for Head Start, Currie and Thomas (1995) use data on children born through 1987 in the CNLSY to show that white children who attended Head Start performed better on the Peabody Picture Vocabulary Test (PPVT) and were less likely to have repeated a grade than their siblings who had not attended. While this was not the case for Black children, Black and white children alike who attended Head Start were more likely to have been immunized against the measles but were no taller around age 5 than their non-attending siblings. In another paper also applying a sibling comparison to the CNLSY, Currie and Thomas (1999) find improvements in PPVT and PIAT scores and reductions in the likelihood of grade retention for Hispanic children who attend Head Start. They also uncover substantial heterogeneity in impacts within this population, with typically larger effects for native-born and Mexican children.

An important contribution of Currie and Thomas (1995) and Currie and Thomas (1999) was in exploring racial and ethnic differences in Head Start effects in nationally representative data. To that point, Head Start research – and research on ECE in general – had been heavily

focused on Black children; indeed, all PPP participants were Black, as were nearly all ABC participants. However, given the timing of the CNLSY and power considerations, only relatively short-term outcomes could be considered.

Deming (2009) was the first to address this limitation. He also applies FFE to the CNLSY, but he can include both more children (later births of the NLSY79 mothers and cohorts born 1976 to 1986) as well as outcomes later in life. Because the PIAT was administered to CNLSY respondents between the ages of 5 and 14, for example, he can trace out how the test score impacts of Head Start evolve as children age, and because a substantial number of CNLSY respondents were old enough to have completed high school, he can consider outcomes in young adulthood. Mirroring findings from the model evaluation literature, he finds test score fadeout, particularly among Black children and disadvantaged children. [21] However, he also finds a significant positive impact of Head Start attendance on an index of young adult outcomes – summarizing impacts on high school graduation, college attendance, idleness, crime, teen parenthood, and health – amounting to a statistically significant 0.23 s.d. increase.

Deming (2009) was the first quasi-experimental study of Head Start to consider both short- and medium-term outcomes. Unlike previous work, Deming (2009) also presents evidence consistent with the idea that, within families, Head Start attendance is randomly assigned. Bauer and Schanzenbach (2016) take the FFE approach to slightly more recent cohorts (born through 1990) and yet more outcomes. They find that Head Start participation has positive effects on self-control and self-esteem, particularly for those whose mothers did not have a high school degree and Black participants. They also see that Head Start participants engage in more positive

[21] On average, impacts fade from a statistically significant 0.14 s.d. at ages 5 to 6 and 7 to 10 to an insignificant to 0.055 s.d. at ages 11 to 14. This overall pattern of effects is driven by Blacks, for whom PIAT score impacts fade from 0.29 s.d. (ages 5 to 6), to 0.13 s.d. (at ages 7 to 10) to 0.03 s.d. (at ages 11 to 14).

parenting practices than their siblings who did not attend Head Start. Like the findings in Barr and Gibbs (2019), this result suggests that Head Start may have intergenerational impacts.

That said, Pages et al. (2020) recently update Deming (2009) by considering even later-life outcomes, like earnings, for Deming's original cohorts, as well as more recent cohorts, born into the 1990s (1987 to 1996). Applying the same FFE approach, they find a smaller positive impact on a summary index of adult outcomes for the same cohorts considered by Deming (2009) (0.17 s.d.). Years of education was the only outcome in the index for which there was a significant positive effect (0.3 years); impacts on earnings and college attendance were small and not statistically significant. For the more recent cohorts, moreover, siblings who attended Head Start scored worse on the summary index of adult outcomes. The authors' preferred explanation is that more recent cohorts in the CNLSY have older mothers and come from better-resourced families, making their counterfactual to Head Start higher quality than would have been experienced by earlier cohorts, born to younger mothers.

Carneiro and Ginja (2014) take the RDD to the CNLSY, comparing children whose family incomes would have put them in the neighborhood of being eligible for Head Start between the ages of 3 and 5, focusing on cohorts born 1977 to 1996. While they find no first-stage impact of income eligibility on Head Start attendance for girls, that impact is substantial for boys. So, too, are the ITT impacts of eligibility on summary indices of the behavioral, health, and cognitive outcomes of boys at ages 12-13 and 16-17, with noteworthy reductions in the likelihood of overweight at both ages and in depression in the older age group. Male Head Start participants are less likely to have been convicted of a crime or arrested by age 20-21 and are less likely to be idle. While Carneiro and Ginja (2014) do not present separate estimates by cohort, they do note that the inclusion of more recent cohorts than Deming (2009) may be one

reason that they see no positive impacts on the PIAT scores of boys at 12 to 13; another possible explanation is that their estimation sample is less disadvantaged.

3. Head Start Impact Study

Head Start appears to have delivered more positive longer-term outcomes for cohorts that participated closer to its inception (see also Duncan and Magnuson (2013)). As alluded, one possible explanation is that the learning environments experienced in the absence of Head Start – the counterfactual – may have been higher quality for more recent cohorts. In the CNLSY, this could be in part an artifact of data construction. Mothers of all children in the CNLSY were born between 1957 and 1964. More recent cohorts are thus born to older mothers, whose lives (and livelihoods) are more established, allowing them to provide better care at home or to purchase private center-based care. But this finding should also hold in other data: The counterfactual to Head Start participation has become increasingly likely to be another center-based program, given the spread of pre-K programs and the rise of maternal employment since the 1960s.

The HSIS has provided not just the first experimental evidence on Head Start, but also a unique window into the importance of the counterfactual *within* recent cohorts. Congressionally mandated as part of Head Start's 1998 re-authorization, the HSIS randomized Head Start offers among almost 5000 3- and 4-year-olds at over-subscribed Head Start centers across the country in fall 2002 (i.e., across children born in the late 1990s). The goal was to determine "the impact of Head Start on children's school readiness, and parental practices that support children's development" and "under what circumstances Head Start achieves its greatest impact" (Puma et al., 2010). Study participants were followed through third grade. Internal evaluation of the experimental data by Puma et al. (2010) found that children randomly assigned into the treatment

group initially had better cognitive outcomes than their counterparts in the control group. However, these differences rapidly faded.[22]

Others have replicated and extended these results. Kline and Walters (2016) focus on the literacy and math components of the WJAT observed consistently over time and on estimation of the TOT/LATE, instrumenting for Head Start attendance with the randomized offer conditional on site fixed effects using 2SLS. They also present estimates that pool across the 3- and 4-year-old cohorts. Magnitudes differ, but the basic pattern of findings is consistent with those from the internal evaluation: Head Start attendance raises test scores initially, but these effects fade rapidly. Still, their cost-benefit analysis, which follows Chetty et al. (2011) in assuming that earnings gains from early intervention can be predicted from initial test score gains, suggests the benefits exceed the costs. The benefit-cost ratio becomes more favorable when the fiscal externalities from program substitution are taken into account.

Instead of focusing on mean impacts with a standard linear regression analysis, Bitler, Hoynes, and Domina (2014) estimate the effects of Head Start across the distribution of test scores using quantile regression. Accounting for non-compliance with offer assignment using an instrumental variables approach, the authors find that participation raises test performance by more at the bottom of the distribution than at the middle or the top. They also show that ITT impacts are not systematically larger for population subgroups more likely to have been exposed to informal or maternal care in the absence of the intervention. One interpretation is that the counterfactual care experience is not an important mediator of program effects.

Such an interpretation is, however, at odds with two later analyses of how the counterfactual influences the magnitude of treatment effects in the HSIS. First, Kline and

[22] Puma et al. (2010) focus on presentation of ITT estimates, though in an appendix provide TOT/LATE estimates, which account for non-compliance in the experiment offers.

Walters (2016) begin with the observation that the "subLATEs" for two key subgroups – those who would have otherwise been in a center-based care situation and those who would have otherwise been at home or in informal care – can in principle be identified from a linear model with two endogenous variables – a dummy for Head Start attendance and a dummy for other preschool attendance.[23] Identifying such a model requires an instrument (or instruments) for both. They consider several sets of instruments, ultimately finding that interactions between the (random) Head Start offer and both covariates and indicators for six site types (grouped based on program substitution patterns) produce reasonably strong first stages. Though the findings suggest that both Head Start attendance and other preschool attendance raise test scores relative to home or informal care at the end of the Head Start year, the overidentification test rejects the model.

Kline and Walters (2016) therefore move to a choice model that allows for selection on unobserved characteristics. To identify the effects of attendance in each program type, they estimate this model, including multiple control functions (akin to inverse Mills ratios in the standard Heckman selection model) that incorporate the instruments described above. The model estimates imply that the subLATE for the two-thirds of children who would have otherwise been at home or in informal care ("home compliers") is a significant 0.37 s.d.; for the remaining third of children who would have otherwise been in other center-based care ("center compliers"), the subLATE estimate is -0.1 s.d., but not statistically significant.[24]

[23] There is a parallel here to the FFE literature on Head Start. Papers in this space also include an indicator for other preschool attendance. If Head Start attendance and other preschool attendance are randomly assigned within families, the coefficient on the Head Start indicator thus identifies the "subATE" for an individual who would have otherwise been at home or in informal care, and the difference between the Head Start and other preschool coefficients identifies the "subATE" for those who would have otherwise been in another preschool.

[24] A nice feature of the Kline and Walters (2016) approach is that it allows the authors to explore the nature of selection into Head Start participation. They find evidence of "a 'reverse Roy' pattern of selection whereby children with unobserved characteristics that make them less likely to enroll in Head Start experience larger test score gains"

Feller et al. (2016) come to a substantively similar conclusion using a Bayesian principal stratification framework (Frangakis and Rubin, 2002). Focusing on the PPVT rather than WJAT scores due to limitations imposed by this particular method, they find subLATEs of 0.23 and 0.00 s.d. for home compliers and center compliers, respectively, at the end of the Head Start year. They also explored heterogeneity in these subLATEs across groups defined on the basis of pre-existing observables, finding relatively large impacts for home compliers who scored in the bottom third of the pre-intervention test (versus not) and who were dual language learners (versus not). Together with the findings from Kline and Walters (2016), these results suggest that Head Start continues to deliver moderate-sized short-term test score impacts relative to home or informal care, much as was the case for the cohorts studied in Deming (2009).[25]

C. State-Funded Kindergarten and Pre-K

Head Start is, of course, not the oldest public early education program in the U.S. Kindergartens, for 5-year-old children, came first. However, kindergartens were not universally provided by school districts across the country until the 1980s, as discussed in Section II. While their foothold in U.S. public education remains much more tenuous, with few pre-K programs in principle serving all children regardless of need, funding of American pre-K programs also did not become mainstream until the 1980s and 1990s.

1. Kindergarten: DD approaches

Cascio (2009a) exploits cross-state variation in the timing of the introduction of state subsidies for universal kindergarten programs in the South in the 1960s through 1980s to estimate the longer-term impacts of kindergarten exposure and attendance. She finds that the

(p. 1798). Cornelissen et al. (2018) also document a reverse Roy pattern of selection in their study of the German universal childcare system.

[25] Zhai, Brooks-Gunn, and Waldfogel (2014) also come to similar conclusions using principal score matching and weighting, which assumes selection on observables only.

introduction of kindergartens led to modest improvements in the later-life outcomes of southern-born white children, but not southern-born Black children. Exploring several potential explanations for this finding, she concludes that southern-born Black children were more likely to have been drawn into public school kindergartens from Head Start, which enrolled a significant number of 5-year-olds when kindergartens were not widespread (Zigler and Meunchow, 1992). The relatively small long-term impacts even for white children suggest that these public-school kindergartens were low-quality relative to the alternative.[26]

The variation exploited in Cascio (2009a) came from the end of more than a century-long kindergarten movement. Indeed, Figure 1 suggests that 5-year-old enrollment rates may have been higher in 1920 than in 1930 or 1940. In a recent working paper, Ager and Cinnirella (2020) combine city panel data on the establishment of kindergartens across the U.S. from 1880 through 1910 with full-count Census data to estimate the impacts of exposure to these early kindergartens. Families with higher exposure experienced larger reductions in fertility, and children with higher exposure were less likely to engage in child labor, stayed in school longer, and had higher occupational prestige and earnings as adults.[27] Effects are larger for immigrants and the disadvantaged. At this time, there were essentially no formal ECE alternatives.

2. Pre-K: A Mix of Approaches

a. Program Implementation: DD

There has been significantly more research on state-funded pre-K programs, some of it also using program introduction as a source of identification. Many of these studies have focused

[26] Dhuey (2011) offers a follow-up study on Cascio (2009a) that incorporates more states, including those with earlier funding initiatives.

[27] In related work, Herbst (2017) estimates the later life impacts of the universal childcare programs established through the U.S. Lanham Act of 1940. Exploiting variation across states and cohorts, he finds that exposure to Lanham Act funding in childhood led to small increases in a summary index of adult outcomes incorporating employment, earnings, and use of public assistance.

on pre-K programs that, like kindergarten, are universal, or available to all children who meet age guidelines where they are offered. [28]

Multiple studies exploit the roll-out of universal pre-K programs in Georgia and Oklahoma – the first two adopting states – for identification. Constructing a synthetic comparison group, Fitzpatrick (2008) finds that the 1995 implementation of universal pre-K in Georgia raised the National Assessment of Educational Progress (NAEP) reading and math scores of fourth-graders residing in rural areas and increased their likelihood of being on grade for age. Building on this result, Cascio and Schanzenbach (2013) present suggestive DD-based evidence that the introduction of the universal pre-K programs in Georgia and later Oklahoma (in 1998) raised reading and math NAEP scores not just in fourth grade, but also math scores as late as eighth grade, though only for low-income (free or reduced-price lunch eligible) students. For higher-income children, there was substantial crowd-out of private preschool.[29]

In a recent working paper, Zerpa (2021) takes advantage of the introduction (or significant expansion) of state-funded pre-K across 10 states between 1998 and 2005 to estimate the short- and medium-term effects of pre-K on non-test score outcomes. She finds evidence of

[28] Several relevant papers do not fit neatly into the categorization of methods in Section III. First, Bartik and Hershbein (2018) take a two-way fixed effects approach to estimate the effects of a cohort's public school pre-K enrollment rates (at the state, district, and school) levels and various outcomes in grades 4 and 8, finding that the average pre-K program does not improve reading or math scores or reduce special education placement or grade retention. Challenges with this approach are that the source of cohort-by-area variation in public school pre-K enrollment rates is unclear, and enrollment rates are likely measured with error, particularly at finer levels of geography, due to the fact that not all state-funded pre-K programs are operated through public schools. (See also Rosinsky (2014)). Second, using data from the 1998-99 kindergarten cohort of the Early Childhood Longitudinal Study (ECLS-K), Magnuson, Ruhm, and Waldfogel (2007) estimate the effects of pre-K participation on outcomes in kindergarten and first grade, controlling for observables using conventional linear regression as well as propensity score methods. They find that pre-K attendance raises test scores but weakens behavioral outcomes in the fall of kindergarten, and the negative behavioral effects persist to the spring of first grade. Since these estimates could be subject to bias from selection on unobservables, the authors explore using state pre-K funding and enrollment as instruments for pre-K attendance. However, the fact that the ECLS-K includes only one cohort makes this effectively a cross-state comparison. (See also Loeb et al. (2007).)

[29] Bassok, Fitzpatrick, and Loeb (2014) examine how the introduction of the universal pre-K programs in Georgia and Oklahoma affected childcare providers, and Bassok (2012) discusses how state-funded pre-K programs have affected the Head Start programs.

significant reductions in grade repetition, with larger effects for low-income children in universal pre-K states. The finding of larger impacts for universal pre-K programs is consistent with Cascio (2020), described below.

Not all studies rely on variation in program adoption at the state level. Across a series of papers, Kenneth Dodge, Helen Ladd, and Clara Muschkin combine variation across counties in the timing and intensity of funding for North Carolina's state pre-K program with rich state administrative data. Though funding was targeted toward low-income children, most classrooms included non-eligible children, potentially making it more like a universal pre-K program in effect. Ladd, Muschkin, and Dodge (2014) find that more pre-K funding directed to a child's birth county when s/he would have been age-eligible for pre-K raised her reading and math test scores in third grade. Likewise, Muschkin, Ladd, and Dodge (2015) find that increases in state pre-K funding reduce the odds of special education placement in third grade, considerably reducing costs to the state. Dodge, Bai, Ladd, and Muschkin (2016) consider both sets of outcomes, as well as grade retention through fifth grade, finding the effects persist.

b. Age-Eligibility: RDD

A nice feature of a DD approach exploiting variation from program implementation is that it allows for estimation of the medium- and long-term impacts of ECE exposure and attendance. However, DD approaches can only be applied where program implementation is dramatic enough to see an effect in aggregate data – typically in places with universal programs. Targeted pre-K programs are nevertheless quite common (Friedman-Krauss et al., 2020).

The age-eligibility RDD has now been widely applied to estimate the short-term effects of both universal and targeted programs. In the pioneering application, Gormley and Gayer (2005) compare the fall 2001 test scores of public school children in Tulsa, Oklahoma who just

barely made the cutoff to enter kindergarten that year – and who are highly likely to have attended pre-K the year prior – to those of children who just barely missed the cutoff, who were therefore just embarking on their pre-K year. They found that Tulsa's universal pre-K program improved cognitive, motor, and language scores, with larger impacts for minorities and children eligible for free- or reduced-price lunch.[30]

Weiland and Yoshikawa (2013) apply a similar RDD approach to evaluate Boston's universal pre-K program. Like the Tulsa program, Boston's program was well-resourced on many dimensions, but research-based curricula and teacher coaching provided additional scaffolding for program effects. They found "moderate-to-large" impacts on tests of language, numeracy, and mathematics, but small impacts on tests of executive functioning and socio-emotional skills. Like Gormley and Gayer (2005), they found some evidence of larger impacts for free- or reduced-price lunch eligible children and nonwhite children.

Several papers offer RDD estimates for a collection of states simultaneously, including many with targeted programs. Wong et al. (2008) present RDD estimates of pre-K eligibility and attendance impacts on for five states – Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia. Like the Tulsa and Boston programs, the pre-K programs in these states are fairly well-resourced by conventional measures of structural quality (i.e., mandated class sizes or staff ratios, teacher education requirements, available support services; see Section V). However, only one (Oklahoma) was universal at the time the data were collected; the remainder targeted enrollment based on economic need or other risk factors. State-specific sample sizes are also typically smaller than in the Tulsa and Boston pre-K evaluations, contributing to imprecision.

[30] In related work, Smith (2016) finds that Black children (but not white children) just barely eligible for Oklahoma's universal pre-K program in its first year were significantly less likely to have been charged with a crime at ages 18-19 than their counterparts just barely ineligible.

The findings are indeed less conclusive than in Gormley and Gayer (2005) and Weiland and Yoshikawa (2013): Though most point estimates are positive, they are statistically significant for the PPVT in only two cases (New Jersey and Oklahoma) and math in only one case (New Jersey). Positive effects on print awareness are more robust and more common. But small sample sizes preclude fruitful exploration of heterogeneity in impacts by race or socio-economic status.

Barnett et al. (2018) estimate RDD impacts for the five states in Wong at al. (2008), as well as for Arkansas, California, and New Mexico – states with different (and sometimes less rigorous) pre-K standards than the original five, but all with targeted programs.[31] While the sample sizes remain small relative to the Tulsa and Boston evaluations, the external validity of the estimates may benefit from the authors' use of data from many cohorts. Consistent with the findings of Wong et al. (2008), the estimates for literacy were almost universally positive and statistically significant, whereas those for language (PPVT) and math (Woodcock-Johnson) were less likely to be significant at the individual state level. Pooling across states, however, weighted average estimates for all tests are statistically significant, with the largest impacts for literacy, followed by math, then language.

There are two broad sets of limitations with these RDD pre-K evaluations. The first, outlined in Lipsey et al. (2015), are largely idiosyncratic to the type of administrative data universally used in these analyses rather than to the design itself. In particular, use of fall test scores has the potential to confound the effects of kindergarten attendance with the (persistent) effects of pre-K. Additionally, focusing on children who enroll in public schools, rather than complete cohorts, precludes identification of proper ITT and TOT/LATE coefficients. Second, despite the relatively large number of contexts where the same basic RDD has been applied and

[31] The West Virginia program became universal across the cohorts studied in Barnett, et al. (2018).

even often the same tests used, little attention has been paid to understanding the sources of cross-state variation in program impacts – for example, whether universal programs have systematically different impacts than targeted ones.

Cascio (2020) tries to address these limitations using data from the Birth Cohort of the Early Childhood Longitudinal Study (ECLS-B), a survey following a nationally representative sample of children born in 2001 from birth through kindergarten completion. Unlike prior RDD studies, the paper harnesses the age-eligibility variation in pre-K enrollment in a specification that compares adjacent school entry cohorts in two groups of states: (1) those with robust pre-K programs in 2005-06, when the 2001 birth cohort would have first aged into pre-K eligibility; and (2) those without pre-K programs in 2005-06 or where pre-K programs were too small or too little differentiated across 3- and 4-year-olds to detect a first-stage impact on pre-K attendance in the ECLS-B. This DD approach has the benefit of allowing for the estimation of true ITT and TOT/LATE effects.[32] In addition, by using the same outcome and applying the same methodology to different groups of states, the paper can compare differences in efficacy between universal and targeted programs.

Cascio (2020) finds that universal pre-K eligibility and attendance increases preschool-age test scores by significantly more than targeted pre-K program eligibility and attendance, with a larger universal-targeted difference in impacts for children eligible for free or reduced-price lunch. Differences in program impacts between universal and targeted pre-K are not explained by systematic differences in other program characteristics (i.e., which standards are applied), differences in state demographics, or differences in the counterfactual care environment. Akin to

[32] The underlying variation is, however, less local to the age-eligibility threshold, potentially raising questions about internal validity. Informal testing of the model's identifying assumption (i.e., balance tests using birth weight and earlier developmental test scores as outcomes) suggest that it holds.

the body of RDD evaluations, there are larger and more robust impacts for reading over math. But as with the age-eligibility RDD, longer-term impacts cannot be considered.

c. Social Experiments

Following the arc of research on Head Start, pre-K research on recent cohorts has had the potential to benefit from social experiments that do not face this limitation. The Tennessee Voluntary Pre-K (TN-VPK) evaluation, conducted by researchers at Peabody Research Institute Vanderbilt University in coordination with the Tennessee Department of Education, randomized offers to attend this highly resourced, targeted program to around 3000 children across oversubscribed centers in the state in fall 2009 and fall 2010. In principle, the TN-VPK evaluation could thus be carried out in the same way as the HSIS evaluation. However, consent rates for follow-up testing that were both low and significantly different across treatment status led the researchers to abandon this approach in favor of a comparison between actual participants and non-participants, adjusted for selection on observables. They find test score impacts at the end of the pre-K year similar in magnitude to those seen in the HSIS, but negative effects as of third grade (Lipsey, Farran, and Durkin, 2018).

More recently, Weiland et al. (2019) present findings from a randomized evaluation of the Boston Public Schools (BPS) Pre-K program, effectuated by the use of lotteries to assign slots in oversubscribed programs between fall 2007 and fall 2011. Although lottery winners were more likely both to attend BPS Pre-K and to remain enrolled in the BPS through third grade than lottery losers, they were not less likely to be retained in grade or to be placed in special education, and their third grade math and reading scores were no higher. The authors show, however, that in the lotteries under study, the marginal BPS pre-K enrollee would have otherwise

received other formal care or education. The null findings are thus consistent with the HSIS analyses of Feller et al. (2016) and Kline and Walters (2016).

Thus, despite their promise, these two social experiments face obstacles in providing credible estimates of the longer-term impacts of targeted and universal pre-K. Adding rigor to the TN-VPK evaluation will require administrative data on the full study sample. In addition, the essentially complete crowd-out of center-based care in the BPS Pre-K evaluation means that longer-term follow-ups are likely to deliver null findings – or at least will not be able to say anything about the subLATE for pre-K "home compliers," who are still likely to be substantial in many settings. There is thus still much to be learned about the impacts of pre-K, particularly concerning longer-term effects.

V. The Importance of the Counterfactual and Program Quality

Several common themes emerge from this literature review. First, any given ECE program appears to have smaller impacts on outcomes when it substitutes for another. Second, although not always the case, the same program tends to have larger impacts on those who are more disadvantaged, even when the scope for program substitution is limited. These observations led Cascio and Schanzenbach (2014) to develop a simple framework to summarize existing literature and to guide interpretation of prospective studies. The accompanying schematic illustrates the importance of the counterfactual but also reveals how little the literature has addressed what defines quality ECE in practice. I end this section with a brief discussion of the literature on ECE quality, or the "ECE production function."

A. A Framework for Interpretation

Suppose that (1) The higher the quality of a child's learning environment – accounting for both time at home and school or childcare – the more a young child will learn over any

period of time;[33] and (2) In the absence of ECE intervention, the quality of a child's learning environment is (weakly) increasing in family socio-economic status (SES).[34] An implication of assumption (1) is that ECE treatment effects should be directly proportional to the change in the quality of the learning environment that the ECE program in question represents, relative to the counterfactual. An implication of assumption (2) is that an ECE program of the same quality will generate weakly larger ECE treatment effects for those from lower-SES backgrounds.

Cascio and Schanzenbach (2014) illustrate the implications through a simple schematic. Figure 6 Panel A considers a benchmark case, where the quality of the learning environment is linearly increasing in SES in the absence of formal ECE. The slope of this solid bold line could reflect the SES gradient in early assessment scores under the counterfactual, assumed here to be exclusive maternal or informal care. The remaining horizontal lines then reflect the canonical understanding of PPP and Head Start at its inception, when such a counterfactual would have been more likely to hold for targeted children due to lower rates of maternal labor force participation and a less developed private care and education market. Reflecting the findings on these programs outlined in Section IV – and their common interpretation – I have modeled the vertical distance between the quality of PPP and the quality of the counterfactual to be longer, and thus PPP treatment effects to be larger. The distance is thus shorter for Head Start both because it casts a "wider net," reaching further up the distribution of SES, but also because Head Start quality is widely considered lower than that in PPP.

In more recent years, the counterfactual learning environment for the average ECE participant has not been one with exclusive maternal or informal care. For at least some children

[33] I will defer discussion of how exactly quality can be measured; for now, just assume that quality is defined such that this assumption holds.
[34] Higher SES parents may be better teachers themselves and have more resources to invest in their children's human capital.

today, a randomized offer to attend Head Start or establishment of a preschool in the local public schools displaces attendance in formal childcare or private preschool. As described in Section IV, the ITT and TOT/LATE then become weighted averages of ITTs and TOTs/LATEs for children who would have been in a formal learning situation and children who would have been in informal or home care, with weights proportional to each group's share in the population. Assuming that a formal learning situation is superior to maternal or informal care, the quality of the counterfactual learning environment then increases for the average child. Figure 6 Panel B models other center-based care as having a larger impact on lower-SES children.

The consequence is to erode estimated treatment effects relative to what they would have been in the absence of formal learning opportunities. In Figure 6 Panel B, I depict the two valid U.S. social experiments discussed in Section IV. I assume that Head Start, as it pertains to the HSIS, served the same range of the SES distribution at roughly the same quality as in Panel A. But now it generates a smaller impact: The vertical distance between Head Start quality and the quality of the average counterfactual learning environment is now lower because it averages across children who would have attended other center-based care (null impacts) and those who would not have (positive effects) (Kline and Walters, 2016; Feller et al., 2016; Zhai, Brooks-Gunn, and Waldfogel, 2014). The other horizontal line in Panel B represents the universal pre-K program in BPS, which is widely regarded as high quality and serves children from across the SES distribution. The Weiland et al. (2019) finding of null effects reflects essentially complete formal program substitution on the margin.

Figure 6 Panel C considers another case for universal pre-K, brought to light by studies that present separate estimates by family background (e.g., Cascio and Schanzenbach, 2013). In this version of the figure, I have added a line representing the quality of the counterfactual

environment for the *average* child by SES, which is a weighted average of the quality of the formal and informal/maternal counterfactuals (now represented with dashed lines); weights on the formal counterfactual are equal to the share experiencing it, which I assume to be increasing in SES, per the statistics in Table 1. Lower-SES children on average continue to have a lower quality counterfactual learning environment both because they are less likely to be enrolled in another program and because of a lower-quality home learning environment. As in the first example, effects are thus larger for lower-SES children. However, they are smaller than they would have been in the absence of formal alternatives.[35]

B. What is Quality?

In this simple framework, program quality is revealed by the magnitude of its impact on outcomes, conditional on the counterfactual. Conceptualizing quality in this way can be misleading. Two programs that have different impacts – like Head Start at its inception and PPP – may not really be that different in terms of quality, but just have different counterfactuals. Likewise, two programs that yield the same-sized (positive) impacts might be very different quality in absolute terms, depending on the counterfactual. In addition, any given program employs a bundle of inputs, combining them in a potentially unique way. From a practical perspective, concerned with resource allocation, it may be more helpful to understand which inputs and production technologies matter.

ECE researchers outside of economics have categorized ECE quality into two types – *structural quality* and *process quality*. According to Yoshikawa et al. (2013):

> Process quality features—children's immediate experience of positive and stimulating interactions—are the most important contributors to children's gains in language, literacy, mathematics, and social skills. Structural features of quality (those features of quality that can be changed by structuring the setting differently

[35] For the highest SES children, treatment effects can even be negative. Havnes and Mogstad (2015) find negative effects in the top tercile of the income distribution in a DD analysis of universal childcare in Norway.

> or putting different requirements for staff in place, like group size, ratio, or teacher qualifications) help to create the conditions for positive process quality, but do not ensure that it will occur (p. 6).

In the language of the economics of the education, structural quality is thus embodied in inputs, like class size and teacher qualification requirements, that can be relatively readily manipulated through policy intervention. By contrast, process quality is teacher quality; some teachers are better than others at providing the positive and supportive interactions that stimulate early learning with the inputs available. The closest analogy to process quality in the economics of education is thus teacher value-added (see, for example, Chetty, Friedman, and Rockoff (2014) and Jackson (2018)).

Estimation of "education production functions" for elementary and secondary education, relating school inputs to school outputs for attendees, has by now been the subject of a large literature (see, for example, chapters by Glewwe (for developing countries) and Chakrabarti (for developed countries) in this *Handbook*), as has teacher value-added (see the chapter by Springer in this *Handbook*). At present, we have much less evidence on "ECE production functions," and very little evidence on the value-added of ECE teachers. Generating such evidence faces similar challenges as the literature on attendance and the parallel quality and teacher value-added literature in K-12 education: ECE quality and teachers are typically not randomly assigned. However, it also faces unique challenges; for example, estimation of teacher value-added models in the ECE context is often prohibitively expensive, since most young children are not yet literate and are thus incapable of being assessed without one-on-one adult supervision.

Given these challenges, the most convincing evidence on ECE production functions to date has come from a limited number of studies that either randomly assign some aspect of ECE quality (like class size or length of the school day) or randomly assign students to ECE sites (or

within site, to teachers), and provide contemporaneous (with the intervention) testing data on all study participants. The remainder of this section discusses the limited literature.

C. Evidence on Quality in ECE

The first studies that meet these criteria came from the Tennessee STAR (Student-Teacher Achievement Ratio) experiment, or "Project STAR," which randomly assigned about 6000 entering kindergartners into different-sized classes within 79 Tennessee schools in the fall of 1985. New school entrants over the next three years were randomly assigned to classrooms, and all study children maintained their experimental class size types through third grade. Re-evaluating the initial experimental data, Krueger (1999) finds that children randomly assigned to smaller classes, particularly minority and lower-income children, had higher test scores through third grade. Follow-up studies on Project STAR have shown that that the class size intervention boosted test scores in middle school and increased the chances of taking a college-entrance exam and completing college, particularly for minorities (Krueger and Whitmore, 2001; Dynarski, Hyman, and Schanzenbach, 2013). Merging Project STAR participants to tax records, Chetty et al. (2011) likewise find that small classes raise college attendance, though not earnings at age 27.

Because Project STAR also randomly assigned *teachers* to classrooms within schools, teacher characteristics were orthogonal to both class size and student characteristics. While Krueger (1999) finds that teacher characteristics like education and experience did not affect test scores in the short term on average, Dee (2004) shows that being assigned a teacher of the same race significantly increased them, and Chetty et al. (2011) find that children assigned more experienced kindergarten teachers had higher earnings at age 27. Project STAR's design also means that it has been possible to identify classroom effects. Chetty et al. (2011) found that

children randomly assigned to classmates scoring higher on end-of-year exams were themselves better off later in life – more likely have attended college and higher earning.

More innovation has occurred outside of the U.S. context. Araujo et al. (2016) explicitly study the importance of teacher quality in kindergarten, randomly assigning nearly 25,000 children across two cohorts to teachers within Ecuador elementary schools. Their data allow them to produce experimental estimates of teacher value-added and to correlate teacher value-added against a number of teacher characteristics – not just education and experience, but also teacher family background, teacher personality characteristics, and scores on an observational rubric of child-teacher interactions implemented by trained assessors – the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, and Hamre, 2008).[36] CLASS is widely used in ECE contexts and has been found to be more predictive of children's learning gains in ECE than structural quality (Sabol et al., 2013). However, until Araujo et al. (2016), it had not been directly compared to teacher test score value-added. Araujo et al. (2016) find that significant variation in efficacy across teachers and that CLASS scores correlate highly with teacher efficacy in math, reading, and a measure of executive function.

Returning to the U.S. context, Bloom and Weiland (2015) and Walters (2015) estimate significant heterogeneity in the impacts of Head Start attendance across sites in the HSIS. Bloom and Weiland (2015) consider both cognitive and socio-emotional outcomes and present estimates of cross-site heterogeneity in the treatment-control contrast in care experiences (e.g., in terms of average weekly hours in center care, percent with a teacher who has a BA). However, Walters (2015) formalizes the connection, estimating the correlation between various characteristics of

[36] Araujo, Dormal, and Schady (2019) examine how caregiver characteristics relate to developmental outcomes of infants and toddlers in a setting with within-center random assignment to caregivers. They find that children randomly assigned to teachers with higher CLASS scores or more experience had better development outcomes; however, caregiver education—a structural quality measure—has no impact.

the local environment and site-specific treatment effects. He finds that Head Start sites where there is a greater substitution from center-based ECE tend to exhibit lower effects on cognitive test scores, consistent with findings from the HSIS earlier reported. Teacher education, the High/Scope (PPP) curriculum, and class size did not correlate with test score impacts, but Head Start centers with full-day programs and home visitation tended to have larger test score effects.

A limitation of the Walters (2015) analysis is that, although the HSIS randomized offers to attend Head Start within each participating site, it did not randomly assign local program characteristics. Several recent studies overcome this limitation in the context of estimating impacts of the length of the school day. Gibbs (2014) uses lottery-based variation from oversubscribed programs to estimate the effects of full-day (relative to half-day) kindergarten in five Indiana school districts, finding that full-day programs increase literacy scores, particularly among Latino children and children with low baseline literacy levels. More recently, Atteberry, Bassok, and Wong (2019) carry out a similar analysis in the context of pre-K in a school district near Denver, Colorado. They find that full-day pre-K increases receptive vocabulary skills and teacher-reports of cognitive, physical, and socio-emotional development in pre-K as well as literacy tests at the start of kindergarten.

VI. Why: Theoretical Mechanisms

Another limitation of the simple framework laid out in Section V is that it does not provide any insight into *why* ECE might work to improve a child's outcomes. Broadly, there are two sets of theoretical mechanisms: those that focus on how program participation itself affects human capital formation and those that focus on how program participation might affect other aspects of a child's environment, which affect human capital formation in turn.

A. Program Participation and Human Capital Formation

1. The Technology of Skill Formation and Dynamic Complementarity

In terms of how program participation itself affects human capital formation, the dominant theory among economists has been articulated by Heckman (2007) and Cunha and Heckman (2007, 2009). This theory concerns the process by which investments are combined to form skill across childhood or the technology of skill formation. Importantly, "skill" in this framework is multi-dimensional, including not just the cognitive (as might be measured on IQ and achievement tests), but also the "non-cognitive" – all else that might contribute to the measures of long-run well-being that have been the focus of the ECE literature. Investments in one period raise the child's skill or capabilities in the next, and different types of skill can complement one another so that the whole is more than the sum of the parts. More skill also encourages future investment by increasing its return. With a stronger foundation through ECE, subsequent investments yield a higher payoff, making a person better off later in life than they would have been in the absence of ECE, or with investment later in childhood.

One implication of this framework is that investments later in childhood or adolescence should be more productive, or generate larger impacts when they follow larger early life investments or build on larger early life endowments – a phenomenon called "dynamic complementarity." To use the notation of Cunha and Heckman (2009), skill or capabilities in period $t+1$, θ_{t+1}, are a function of skill or capabilities in period t, θ_t, investments in t, I_t, and parental capabilities at t, θ_t^P, i.e., $\theta_{t+1} = f_t(\theta_t, I_t, \theta_t^P)$. Period $t+1$ skill or capabilities are increasing in both period t skill and investments, or $\frac{\partial f_t(\theta_t, I_t, \theta_t^P)}{\partial \theta_t} > 0$ and $\frac{\partial f_t(\theta_t, I_t, \theta_t^P)}{\partial I_t} > 0$. Dynamic complementarity implies further that $\frac{\partial^2 f_t(\theta_t, I_t, \theta_t^P)}{\partial \theta_t \partial I_t} > 0$.

While relatively large internal rates of return to some ECE programs motivated this idea, they do not provide a true test of it. A true test requires there to be two independent investment

shocks at different stages of development – in terms of the notation above, one earlier investment, at say $t-j$, I_{t-j}, which serves to raise skill or capabilities at time t, θ_t, and another later investment, I_t. As pointed out by Almond and Mazumder (2013), a true test of dynamic complementarity in observational data in essence thus requires "lightning to strike twice" (p. 49), and it is typically challenging enough to find variation suitable for credible program evaluation of ECE alone. Of the literature reviewed in Section IV, only one study – that by Johnson and Jackson (2019) – meets this criterion, in exploiting variation from the rollout of Head Start (I_{t-j}) and variation from subsequent school spending from school finance reform (I_t). They find evidence consistent with dynamic complementarity: SFR-induced increases in school spending are more productive when preceded by greater exposure to Head Start. But this has not always been the case in settings considering ECE. For example, Rossin-Slater and Wüst (2020) find that a targeted preschool program in Denmark had larger later-life impacts in the absence of exposure to a nurse home visiting program in infancy.

2. Reconciling with Cognitive Test Score Fadeout

As described in Section IV, it is common to find "fadeout" in the cognitive impacts of ECE participation. The multi-dimensionality of skills in the framework described above has provided a convenient way to reconcile fadeout with the fact that ECE can also have positive effects on later-life, non-test outcomes. Even if the impact of ECE on cognitive skills may fade, the impact on the non-cognitive skills that are important determinants of later-life outcomes may persist. For example, Heckman, Pinto, and Savelyev (2013) show that, in the context of the PPP, "personality skills" – rather than cognitive skills – explain a lot of the program's effects on adult outcomes. However, it is generally difficult to isolate causal mechanisms in a setting where the only (quasi) experimental variation is in program participation itself. Without adding structure or

having a second source of credible identifying variation, one can typically only identify whether participation affected an outcome that reflects a hypothesized mechanism (e.g., longer-term non-cognitive skills), not also the effect of that intermediate outcome on the later-life outcomes of interest.

Problems with test score measurement may provide another way to reconcile fadeout with persistent effects on later-life, non-test outcomes. In particular, fadeout could be a statistical artifact – an outgrowth of the common practice of re-expressing test scores in distributional terms, such as in standard deviation units. The hypothesis, first articulated by Lang (2010), is that if the distribution of knowledge becomes higher variance as children age, the impacts of an early intervention on standardized test scores will fall even as the effects on knowledge remain unchanged. Cascio and Staiger (2012) test this hypothesis using estimates of the parameters of a model of knowledge accumulation across the school career to predict how the distribution of knowledge expands with child age. Their results imply that the distribution of knowledge does widen as children progress through school, but not by enough to account for much fadeout. Investigating this question from a different angle, Wan et al. (2021) succeed in eliminating fadeout in math scores in an RCT of an early mathematics curriculum intervention with some order-preserving transformations of the original test scale.

These findings suggest that at least some fadeout is likely real, generated by children forgetting what they learned, or by later schooling experiences not reinforcing or building on early knowledge gains. Subsequent attendance at relatively low quality schools is a candidate explanation for stronger test score fadeout among Black Head Start participants (Currie and Thomas, 2000).

B. Other Channels Affecting Human Capital Formation

The mechanisms described in the prior section implicitly assume that the instruction and care provided in an ECE program itself generate program effects. But ECE is not just about the child's experience in the program; other aspects of a child's environment – whether his mother is employed, parenting practices – may change as a result, feeding back into his development.

As discussed in Section II, U.S. ECE programs have typically been explicitly focused on education or enhancing child development, not on providing subsidized childcare. However, public kindergarten and pre-K as well as Head Start provide an implicit childcare subsidy for the length of the school day, opening the possibility that they affect maternal labor supply. Empirical findings on this front nevertheless suggest that maternal employment is not a strong candidate explanation for the developmental effects of ECE participation in the U.S.

Gelbach (2002) was the first to take a quasi-experimental approach to estimate how the childcare subsidy implicit in public early education affects the labor supply of American mothers. In the spirit of the age-eligibility RDD, he uses quarter of birth dummies as instruments for whether a five-year-old is enrolled in public kindergarten in the 1980 Census. He finds modest impacts on various labor supply measures of married women and slightly larger impacts for single women, albeit only those without another child under the age of five. Fitzpatrick (2012) revisits these findings by applying an age-eligibility RDD to restricted-use data from the 2000 Census containing exact day of birth. Unlike Gelbach (2002), she finds positive impacts on the labor supply of single women with no children under age 5. The difference in findings may result in part from differences in research design, but differences in the context for maternal employment in 2000 versus 1980 likely matter as well.

Cascio (2009b) arrives at similar conclusions as Fitzpatrick (2012) but using variation from the introduction of state subsidies for kindergarten across the South and West. Combining

data from the 1950 through 1990 Censuses and taking both a DD approach (exploiting cross-state variation in timing) and a triple-difference (DDD) approach (exploiting both the variation in timing and the fact that American kindergartens subsidize the care of children only at age 5), she finds that the public school enrollment of a 5-year-old increases the employment of single women without any younger children, but has no impact on the labor supply of other mothers, despite substantial substitution of public for private kindergarten.

Researchers have also applied both DD and RDD approaches to estimating how pre-K enrollment affects maternal employment. The most convincing evidence comes from Fitzpatrick (2010), who uses the RDD and restricted-use 2000 Census data containing exact day of birth to estimate the maternal labor supply impacts of universal pre-K eligibility and enrollment in Georgia and Oklahoma. Four-year-olds just barely eligible for these programs were more likely to be enrolled in preschool, but their mothers were typically no more likely to be working. Using a DD approach taking advantage of the introduction of the Georgia and Oklahoma programs, Cascio and Schanzenbach (2013) find a similar result.[37]

There are also some studies on the maternal labor supply impacts of targeted ECE in the U.S., particularly Head Start. Several studies (Sabol and Chase-Lansdale, 2015; Schiman, 2019; Wikle and Wilson, 2020) attempt to exploit the experimental variation from the HSIS, but arrive at different conclusions, arguably due to how they define estimation samples. Wikle and Wilson (2020) combine their HSIS evaluation with a quasi-experimental analysis exploiting variation over time and across states in the intensity of Head Start funding expansions in the 1990s, along with the fact that Head Start largely serves three- and four-year-old children. Their findings

[37] Sall (2014) takes a two-way fixed effects approach estimating the effects of district-level provision of pre-K across ten southern states, nine of which operated targeted programs over the period of study (1990 to 2006). He finds that district-level pre-K provision raises the labor force participation and employment rates of mothers of four-year-olds who have no younger children, but with stronger effects among married than single women.

suggest Head Start increases employment, but only among single mothers. However, applying Ludwig and Miller's (2007) RDD variation to restricted Census data, Pihl (2020) finds that, at its inception, Head Start reduced employment among single mothers in both the short term and over the longer run.

Broadly, the findings from these studies are consistent with Lubotsky and Qureshi (2018) and Cascio (2017), who show that changes in the employment of mothers whose children are just aging into public preschool school eligibility are small relative to those experienced across motherhood overall. Why doesn't public ECE in the U.S. generate a larger maternal labor supply response? Cascio (2015) notes that the developmental impacts of universal ECE programs across the world tend to be inversely related to their maternal employment impacts, potentially due to differences in program emphases and goals.[38] One way in which these differences are manifest is that the school day rarely aligns with the workday. Several studies outside the U.S. (Martínez A. and Perticará, 2017; Duchini and Van Effenterre, 2020) suggest that the length of the school day can be an important constraint on maternal employment.[39] Another possible explanation lies in

[38] The extensive literatures on the universal childcare programs in Norway and Canada help to illustrate this point. The centerpiece of Quebec's Family Policy, introduced in 1997, was the implementation of full-day kindergarten as well as heavily subsidized universal childcare (at a cost of $5 per day), initially for 4-year-olds but expanding to younger children in the years to follow. Using the rest of Canada as a comparison group for Quebec in an initial comprehensive DD evaluation, Baker, Gruber, and Milligan (2008) find that Quebec's significantly (and substantially) increased employment of married women, but reduced child well-being (health and non-cognitive skills). (See also Lefebvre and Merrigan (2008) and Lefebvre, Merrigan, and Verstaete (2009) on employment impacts.) These effects moreover persisted through early adulthood, manifesting in lower self-reported health and increases in criminal behavior (Baker, Gruber, and Milligan, 2019). By contrast, DD estimates of the impacts of the Norwegian universal childcare program, exploiting municipality-by-time variation in program introduction, show limited impacts maternal labor supply (Havnes and Mogstad, 2011a) but improvements in later life outcomes, such as earnings (Havnes and Mogstad, 2011b) that were concentrated in two bottom terciles of the income distribution (Havnes and Mogstad, 2015). Norway's program met many of the standards of preschool programs; Quebec's did not. See Cascio (2015) for further discussion of universal programs elsewhere in the world.

[39] Randomizing slots in oversubscribed after school programs in a subset of Santiago primary schools, Martínez A. and Perticará (2017) show that expansion of access to after-school care opportunities increases the labor force participation and employment of women who have at least one child not yet eligible for school. Using a DD approach comparing women based on their children's age eligibility, Duchini and Van Effenterre (2020) find that a 2013 reform to French primary schools that added class time on Wednesdays increased maternal monthly wages weekly days worked and reduced the gap in part time work between mothers and fathers.

the lack of maternal supports and subsidized childcare in the U.S. at earlier ages. Most American mothers will have decided whether or not to return to work before a child reaches an age to be eligible for public ECE. Pihl (2020) also offers a novel insight: Public ECE does more than just provide an implicit childcare subsidy. The author argues, for example, that Head Start's facilitation of participation in safety net programs, like cash welfare, and emphasis on parental involvement may weaken single women's attachment to the labor force.

Indeed, ECE programs can affect parenting practices. Parenting may improve as a result of a child's ECE participation, or it may deteriorate. Gelber and Isen (2013) consider parenting through the lens of time use. Using data and variation from the HSIS, they find that parents of children randomized slots in Head Start spent more quality time with their children, doing things like reading or math activities. Cascio and Schanzenbach (2013) find a larger gap between less-educated mothers of 4-year-olds and less-educated mothers of 5-year-olds in the amount of time spent caring for or helping children in states with universal pre-K programs (Georgia and Oklahoma) versus elsewhere in the country. A similar DD is not observed for more-educated mothers. However, their estimates were considerably noisier than those of Gelber and Isen (2013).

VII. Conclusions

This chapter has provided an overview of the "what, when, where, who, how, and why" of ECE in the U.S. Over the past two decades, we have made considerable progress toward understanding the impacts of participation in ECE programs operating at scale in the U.S. (and abroad), as the literature has moved away from – despite being still strongly influenced by – small-scale model ECE interventions like PPP and ABC. And yet, there is still much more to learn, even on the participation margin. For example, the literature would benefit from more

social experiments like the HSIS, particularly in the context of pre-K. At present, we also have no long-term outcomes data from social experiments and are instead predicting impacts on longer-term outcomes, like earnings, from initial test score impacts. In general, there is a dearth of studies on the longer-term effects of pre-K, regardless of research design.

There is, at present, even more uncertainty about the mechanisms linking large-scale ECE interventions to later-life well-being, as well as about the ECE production function. To convincingly study the first issue requires not just "as good as random" variation in ECE participation, but also credible identifying variation in the proposed mechanism itself. To design a study that can do both would be challenging, but not entirely outside the realm of possibility for a field experiment. Convincingly studying the second issue requires a shift in mindset, or an application of methods like those described in Section III toward the ECE "quality" margin, holding participation constant. I hope to see considerable progress on these issues in the decades to come.

References

Ager, Philipp and Francesco Cinnirella. 2020. "Froebel's Gifts: How the Kindergarten Movement Changed the American Family." Mimeo. https://131b71d2-d72e-a4b8-5bd5-ef4ebb97577e.filesusr.com/ugd/eacb99_3ad24d4a45dd49809c845e07ce4c7fff.pdf

Almond, Douglas and Janet Currie. 2011. "Human Capital Development before Age Five." In *Handbook of Labor Economics: Volume 4B* (eds. David Card and Orley Ashenfelter), pp. 1315-1486.

Almond, Douglas, Janet Currie, and Valentina Duque. 2018. "Childhood Circumstances and Adult Outcomes: Act II." *Journal of Economic Literature* 56(4): 1360-1446. http://doi.org/10.1257/jel.20171164.

Almond, Douglas and Bhashkar Mazumder. 2013. "Fetal Origins and Parental Responses." *Annual Review of Economics* 5: 37-56. https://doi.org/10.1146/annurev-economics-082912-110145.

Anders, John, Andrew Barr, and Alex Smith. 2019. "The Effect of Early Childhood Education on Adult Criminality: Evidence from the 1960s through the 1990s." Mimeo. http://people.tamu.edu/~abarr/main2d_12_9_2019.pdf

Anderson, Michael. 2008. "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* 103(484): 1481-1495. https://doi.org/10.1198/016214508000000841

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444-455. https://doi.org/10.2307/2291629

Araujo, M. Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3): 1415-1453. https://doi.org/10.1093/qje/qjw016

Araujo, M. Caridad, Marta Dormal, and Norbert Schady. 2019. "Child Care Quality and Child Development." *Journal of Human Resources* 54(3): 656-682. https://doi.org/10.3368/jhr.54.3.0217.8572R1

Attanasio, O.P., Fernández, C., Fitzsimons, E.O.A., Grantham-McGregor, S.M., Meghir, C., Rubio-Codina, M., 2014. "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial." *BMJ* 349(g5785). https://doi.org/10.1136/bmj.g5785

Atteberry, Allison, Daphna Bassok, and Vivian C. Wong. 2019. "The Effects of Full-Day Prekindergarten: Experimental Evidence of Impacts on Children's School Readiness."

Educational Evaluation and Policy Analysis 41(4): 537-562. https://doi.org/10.3102/0162373719872197

Bailey, M.J., Sun, S., Timpe, B., 2020. "Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency." NBER Working Paper 28268. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w28268

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2019. "The Long-Run Impacts of a Universal Child Care Program." *American Economic Journal: Economic Policy* 11(3) 1–26. https://doi.org/10.1257/pol.20170603

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2008. "Universal Child Care, Maternal Labor Supply, and Family Well-Being." *Journal of Political Economy* 116(4): 709-745. https://doi.org/10.1086/591908

Barnett, W. Steven, Kwanghee Jung, Allison Friedman-Krauss, Ellen C. Frede, Milagros Nores, Jason T. Hustedt, Carolee Howes, and Marijata Daniel-Echols. 2018. "State Prekindergarten Effects on Early Learning at Kindergarten Entry: An Analysis of Eight State Programs." *AERA Open* 4(2): 1-16. https://doi.org/10.1177/2332858418766291

Barr, Andrew and Chloe R. Gibbs. 2019. "Breaking the Cycle? Intergenerational Effects of an Anti-Poverty Program in Early Childhood." EdWorkingPaper No. 19-141. Providence, RI: Annenberg Institute at Brown University. http://www.edworkingpapers.com/ai19-141

Bartik, Timothy J. and Brad J. Hershbein. 2018. "Pre-K in the Public Schools: Evidence from within U.S. States." Upjohn Institute Working Papers. https://doi.org/10.17848/wp18-285

Bassok, Daphna. 2012. "Competition or Collaboration?: Head Start Enrollment During the Rapid Expansion of State Pre-kindergarten." *Educational Policy* 26(1): 96-116. https://doi.org/10.1177/0895904811428973

Bassok, Daphna, Maria Fitzpatrick, and Susanna Loeb. 2014. "Does state preschool crowd-out private provision? The impact of universal preschool on the childcare sector in Oklahoma and Georgia." *Journal of Urban Economics* 83:18-33. https://doi.org/10.1016/j.jue.2014.07.001

Bauer, Lauren and Diane Whitmore Schanzenbach. 2016. "The Long-Term Impact of the Head Start Program." Washington, D.C.: The Hamilton Project. https://www.hamiltonproject.org/assets/files/long_term_impact_of_head_start_program.pdf

Bauernschuster, Stephen and Martin Schlotter. 2015. "Public child care and mothers' labor supply—Evidence from two quasi-experiments." *Journal of Public Economics* 123: 1–16. https://doi.org/10.1016/j.jpubeco.2014.12.013

Berlinski, Samuel and Sebastian Galiani. 2007. "The effect of a large expansion of pre-primary school facilities on preschool attendance and maternal employment." *Labour Economics* 14: 665-680. https://doi.org/10.1016/j.labeco.2007.01.003

Berlinski, Samuel, Sebastian Galiani, and Paul Gertler. 2009. "The effect of pre-primary education on primary school performance." *Journal of Public Economics* 93: 219–234. https://doi.org/10.1016/j.jpubeco.2008.09.002

Berlinski, Samuel, Sebastian Galiani, and Marco Manacorda. 2008. "Giving children a better start: Preschool attendance and school-age profiles." *Journal of Public Economics* 92: 1416–1440. https://doi.org/10.1016/j.jpubeco.2007.10.007

Berlinski, Samuel, Sebastian Galiani, and Patrick McEwan. 2011. "Preschool and Maternal Labor Market Outcomes: Evidence from a Regression Discontinuity Design." *Economic Development and Cultural Change* 59(2): 313–344. https://doi.org/10.1086/657124

Bitler, Marianne P., Hilary W. Hoynes, and Thurston Domina. 2014. "Experimental Evidence on Distributional Effects of Head Start." NBER Working Paper 20434. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w20434

Blanden, Jo, Emilia Del Bono, Sandra McNally, and Birgitta Rabe. 2016. "Universal Pre-school Education: The Case of Public Funding with Private Provision." *The Economic Journal* 126(May): 682–723. https://doi.org/10.1111/ecoj.12374

Bloom, Howard S. and Christina Weiland. 2015. "Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study." Mimeo, MDRC (March). https://www.mdrc.org/sites/default/files/quantifying_variation_in_head_start.pdf

Bouguen, Adrien, Deon Filmer, Karen Macours, and Sophie Naudeau. 2018. "Preschool and Parental Response in a Second Best World: Evidence from a School Construction Experiment." *Journal of Human Resources* 53(2): 474–512. https://doi.org/10.3368/jhr.53.2.1215-7581R1

Campbell, Frances, Gabriella Conti, James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, and Elizabeth Pungello, and Yi Pan. 2014. "Early childhood investments substantially boost adult health." *Science* 343(6178): 1478-1485. http://doi.org/10.1126/science.1248429

Carneiro, Pedro and Rita Ginja. 2014. "Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start." *American Economic Journal: Economic Policy* 6(4): 135-173. https://doi.org/10.1257/pol.6.4.135

Cascio, Elizabeth U. 2009a. "Do Investments in Universal Early Education Pay Off? Long-term Effects of Introducing Kindergartens into Public Schools." Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w14951

Cascio, Elizabeth U. 2009b. "Maternal Labor Supply and the Introduction of Kindergartens into American Public Schools." *Journal of Human Resources* 44(1): 140–170. https://doi.org/10.1353/jhr/2009.0034

Cascio, Elizabeth U. 2015. "The Promises and Pitfalls of Universal Early Education." *IZA World of Labor* 116. https://wol.iza.org/articles/promises-and-pitfalls-of-universal-early-education

Cascio, Elizabeth U. 2017. "Public Investments in Childcare." Hamilton Project Policy Proposal 2017-14. Washington, D.C.: The Hamilton Project. https://www.hamiltonproject.org/assets/files/public_investments_child_care_cascio.pdf

Cascio, Elizabeth U. 2020. "Does Universal Preschool Hit the Target? Program Access and Preschool Impacts." NBER Working Paper 23215. Cambridge, MA: National Bureau of Economic Research. https://www.nber.org/papers/w23215.pdf

Cascio, Elizabeth U. and Diane Whitmore Schanzenbach. 2013. "The Impacts of Expanding Access to High-Quality Preschool Education" *Brookings Papers on Economic Activity*, Fall 2013, 127-178. https://www.brookings.edu/bpea-articles/the-impacts-of-expanding-access-to-high-quality-preschool-education/

Cascio, Elizabeth U. and Diane Whitmore Schanzenbach. 2014. "Expanding Preschool Access for Disadvantaged Children" Proposal 1 in *Policies to Address Poverty in America*, eds. Melissa S. Kearney and Benjamin H. Harris. Washington, D.C.: The Hamilton Project, 19-28, June 2014. https://www.brookings.edu/wp-content/uploads/2016/06/expand_preschool_access_cascio_schanzebach.pdf

Cascio, Elizabeth U. and Douglas Staiger. 2012. "Knowledge, Tests, and Fadeout in Educational Interventions." NBER Working Paper 18038. Cambridge, MA: National Bureau of Economic Research. https://www.nber.org/papers/w18038.pdf

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660. https://doi.org/10.1093/qje/qjr041

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value -Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633-2679. https://doi.org/10.1257/aer.104.9.2633

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schoenberg. 2018. "Who Benefits from Universal Childcare? Estimating Marginal Returns to Early Child Care Attendance." *Journal of Political Economy* 126(6): 2356-2409. https://doi.org/10.1086/699979

Cunha, Flavio and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97(2): 31-47. https://doi.org/10.1257/aer.97.2.31

Cunha, Flavio and James Heckman. 2008. "Formulating, Identifying, and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43(4): 738-782. https://doi.org/10.3368/jhr.43.4.738

Cunha, Flavio and James Heckman. 2009. "The Economics and Psychology of Inequality in Human Capital Development." *Journal of the European Economic Association* 7(2-3): 320-364. https://doi.org/10.1162/JEEA.2009.7.2-3.320

Currie, Janet and Duncan Thomas. 1999. "Does Head Start help hispanic children?" *Journal of Public Economics* 74: 235–262. https://doi.org/10.1016/S0047-2727(99)00027-4

Currie, Janet and Duncan Thomas. 2000. "School Quality and the Longer-Term Effects of Head Start." *Journal of Human Resources* 35(4): 755-774. https://doi.org/10.2307/146372

Datta Gupta, Nabanita and Marianne Simonsen. 2010. "Non-cognitive child outcomes and universal high quality child care." *Journal of Public Economics* 94(1-2): 30-43. https://doi.org/10.1016/j.jpubeco.2009.10.001

Dean, Joshua T. and Seema Jayachandran. 2019. "Attending kindergarten improves cognitive but not socioemotional development in India" Mimeo. https://joshuatdean.com/wp-content/uploads/2019/11/HLC_draft_07Nov2019.pdf

Dee, Thomas S. 2004. "Teachers, Race, and Student Achievement in a Randomized Experiment." *Review of Economics and Statistics* 86(1): 195-210. https://doi.org/10.1162/003465304323023750

De Haan, Monique and Edwin Leuven. 2020. "Head Start and the Distribution of Long-Term Education and Labor Market Outcomes." *Journal of Labor Economics* 38(3): 727-765. https://doi.org/10.1086/706090

Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3): 111-134. https://doi.org/10.1257/app.1.3.111

Deming, David and Susan Dynarski. 2008. "The Lengthening of Childhood." *Journal of Economic Perspectives* 22(3): 71-92. https://doi.org/10.1257/jep.22.3.71.

Dhuey, Elizabeth. 2011. "Who Benefits from Kindergarten? Evidence from the Introduction of State Subsidization." *Educational Evaluation and Policy Analysis* 33(1): 3-22. https://doi.org/10.3102/0162373711398125

Dodge, Kenneth A., Yu Bai, Helen F. Ladd, and Clara G. Muschkin. 2016. "Impact of North Carolina's Early Childhood Programs and Policies on Educational Outcomes in Elementary School" *Child Development* 88(3): 996-1014. https://doi.org/10.1111/cdev.12645

Duchini, Emma and Clémentine Van Effenterre. 2020. "School Schedule and the Gender Pay Gap." Mimeo. https://drive.google.com/file/d/1mN7uWBP7okwvqSNwtJrGzIOImnRylAtd/view

Duncan, Greg J. and Katherine Magnuson. 2013. "Investing in Preschool Programs." *Journal of Economic Perspectives* 27(2): 109-32. http://doi.org/10.1257/jep.27.2.109

Dynarski, Susan, Joshua Hyman, and Diane Schanzenbach. 2013. "Experimental Evidence on the Effect of Early Childhood Investments on Postsecondary Attainment and Degree Completion." *Journal of Policy Analysis and Management* 32(4): 692-717. https://doi.org/10.1002/pam.21715

Elango, Sneha, Jorge Luis Garcia, James J. Heckman, Andres Hojman, D. Ermini, M.J. Rados, J. Shea and J.C. Torcasso. 2015. "The internal rate of return and the benefit-cost ratio of the Carolina Abecedarian Project." Mimeo, University of Chicago, Department of Economics.

Elango, Sneha, Jorge Luis Garcia, James J. Heckman, and Andres Hojman. 2016. "Early Childhood Education." In *Economics of Means-Testing Transfer Programs in the United States, Volume 2* (ed. Robert Moffitt), pp. 235-297.

Elder, Todd and Darren Lubotsky. 2009. "Kindergarten Entrance Age and Children's Achievement: Impacts of State Policies, Family Background, and Peers." *Journal of Human Resources* 44(3): 641-683. https://doi.org/10.3368/jhr.44.3.641

Felfe, C., Nollenberger, N., Rodríguez-Planas, N., 2015. "Can't buy mommy's love? Universal childcare and children's long-term cognitive development." *Journal of Population Economics* 28(2): 393–422. https://doi.org/10.1007/s00148-014-0532-x

Feller, Avi, Todd Grindal, Luke Miratrix, and Lindsay Page. 2016. "Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care Type." *The Annals of Applied Statistics* 10(3): 1245-1285. https://doi.org/10.1214/16-AOAS910

Fitzpatrick, Maria D., 2012. "Revising Our Thinking About the Relationship Between Maternal Labor Supply and Preschool." *Journal of Human Resources* 47(3): 583–612. https://doi.org/10.1353/jhr.2012.0026

Fitzpatrick, Maria D. 2010. "Preschoolers Enrolled and Mothers at Work? The Effects of Universal Pre-Kindergarten." *Journal of Labor Economics* 28(1): 51-85. https://doi.org/10.1086/648666

Fitzpatrick, Maria D. 2008. "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement." *The B.E. Journal of Economic Analysis & Policy* 8(1) (Advances), Article 46. https://doi.org/10.2202/1935-1682.1897

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal stratification in causal inference." *Biometrics* 58(1): 21-29. https://doi.org/10.1111/j.0006-341X.2002.00021.x

Friedman-Krauss, Allison H., W. Steven Barnett, Karin A. Garver, Katherine S. Hodges, G.G. Weisenfeld, and Beth Ann Gardiner. 2020. *The State of Preschool 2019: State Preschool Yearbook*. http://nieer.org/wp-content/uploads/2020/07/YB2019_Full_Report.pdf

Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-Term Effects of Head Start." *American Economic Review* 92(4): 999-1012. https://doi.org/10.1257/00028280260344560

Gelbach, Jonah. 2002. "Public Schooling for Young Children and Maternal Labor Supply." *American Economic Review* 92(1): 307-322. https://doi.org/10.1257/000282802760015748.

Gelber, Alexander and Adam Isen. 2013. "Children's Schooling and Parent's Behavior: Evidence from the Head Start Impact Study." *Journal of Public Economics* 101: 25-38. https://doi.org/10.1016/j.jpubeco.2013.02.005

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeerch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344(6187): 998–1001. https://doi.org/10.1126/science.1251178

Gibbs, Chloe. 2014. "Experimental Evidence on Early Intervention: The Impact of Full-day Kindergarten." Mimeo. https://curry.virginia.edu/sites/default/files/files/EdPolicyWorks_files/34_Full_Day_KG_Impact.pdf

Gibbs, Chloe. 2017. "Does full-day kindergarten reduce achievement gaps?" *Focus* 33(2): 17-19. https://www.irp.wisc.edu/publications/focus/pdfs/foc332d3.pdf

Goodman-Bacon, Andrew. 2018. "Difference-in-Differences with Variation in Treatment Timing." NBER Working Paper 25018. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w25018

Gormley, William T. and Ted Gayer. 2005. "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program." *Journal of Human Resources* 40(3): 533-558. https://doi.org/10.3368/jhr.XL.3.533

Grantham-McGregor, Sally, Yin Bun Cheung, Santiago Cueto, Paul Glewwe, Linda Richter, Barbara Strupp, and the International Child Development Steering Group. 2007. "Developmental potential in the first 5 years for children in developing countries." *The Lancet* 369: 60–70. https://doi.org/10.1016/S0140-6736(07)60032-4

Grantham-McGregor, S.M., C.A. Powell, S.P. Walker, and J.H. Himes. 1991. "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study." *The Lancet* 338(8758): 1–5. https://doi.org/10.1016/0140-6736(91)90001-6

Haeck, Catherine, Pierre Lefebvre, and Philip Merrigan. 2015. "Canadian evidence on ten years of universal preschool policies: The good and the bad." *Labour Economics* 36: 137–157. https://doi.org/10.1016/j.labeco.2015.05.002

Havnes, Tarjei and Magne Mogstad. 2015. "Is universal child care leveling the playing field?" *Journal of Public Economics* 127: 100–114. https://doi.org/10.1016/j.jpubeco.2014.04.007

Havnes, T., Mogstad, M., 2011a. "Money for nothing? Universal child care and maternal employment." *Journal of Public Economics* 95(11-12): 1455–1465. https://doi.org/10.1016/j.jpubeco.2011.05.016

Havnes, T., Mogstad, M., 2011b. "No Child Left Behind: Subsidized Child Care and Children's Long-Run Outcomes." *American Economic Journal: Economic Policy* 3(2): 97–129. https://doi.org/10.1257/pol.3.2.97

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The Rate of Return to the High/Scope Perry Preschool Program." *Journal of Public Economics* 94(1-2): 114-128. https://doi.org/10.1016/j.jpubeco.2009.11.001

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103(6): 2052-2086. https://doi.org/10.1257/aer.103.6.2052

Herbst, Chris M. 2017. "Universal Child Care, Maternal Employment, and Children's Long-Run Outcomes: Evidence from the U.S. Lanham Act of 1940." *Journal of Labor Economics* 35(2): 519-564. https://doi.org/10.1086/689478

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945-960. https://www.jstor.org/stable/2289064

Jackson, C. Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes." *Journal of Political Economy* 126(5): https://doi.org/10.1086/699018

Johnson, Rucker C. and C. Kirabo Jackson. 2019. "Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending." *American Economic Journal: Economic Policy* 11(4): 310-349. https://doi.org/10.1257/pol.20180510

Kline, P., Walters, C.R., 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *Quarterly Journal of Economics* 131(4): 1795–1848. https://doi.org/10.1093/qje/qjw027

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497-532. https://doi.org/10.1162/003355399556052

Krueger, Alan B. and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111(January): 1-28. https://doi.org/10.1111/1468-0297.00586

Ladd, Helen F., Clara G. Muschkin, and Kenneth A. Dodge. 2014. "From Birth to School: Early Childhood Initiatives and Third-Grade Outcomes in North Carolina." *Journal of Policy Analysis and Management* 33(1): 162-187. https://doi.org/10.1002/pam.21734

Lang, Kevin. 2010. "Measurement Matters: Perspectives on Education Policy from an Economist and a School Board Member." *Journal of Economic Perspectives* 24(3): 167-182. https://doi.org/10.1257/jep.24.3.167

Lefebvre, Pierre and Philip Merrigan. 2008. "Child-Care Policy and the Labor Supply of Mothers with Young Children: A Natural Experiment from Canada." *Journal of Labor Economics* 26(3): 519–548. https://doi.org/10.1086/587760

Lefebvre, Pierre, Philip Merrigan, and Matthieu Verstraete. 2009. "Dynamic labour supply effects of childcare subsidies: Evidence from a Canadian natural experiment on low-fee universal childcare." *Labour Economics* 16: 490-502. https://doi.org/10.1016/j.labeco.2009.03.003

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* 45: 155-176. https://doi.org/10.1016/j.ecresq.2018.03.005

Lipsey, Mark W., Christina Weiland, Hirokazu Yoshikawa, Sandra Jo Wilson, and Kerry G. Hofer. 2015. "The Prekindergarten Age-Cutoff Regression-Discontinuity Design: Methodological Issues and Implications for Application." *Educational Evaluation and Policy Analysis* 37(3): 296-313. https://doi.org/10.3102/0162373714547266

Loeb, Susanna, Margaret Bridges, Daphna Bassok, Bruce Fuller, and Russell W. Rumberger. "How much is too much? The influence of preschool centers on children's social and cognitive development." *Economics of Education Review* 26: 52-66. https://doi.org/10.1016/j.econedurev.2005.11.005

Lubotsky, Darren and Javaeria Qureshi. 2018. "Assessing the Smooth Rise in Mothers' Employment as Children Age." *Journal of Human Capital* 12(4): 604-639. https://doi.org/10.1086/700077

Ludwig, Jens and Douglas L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics* 122(1): 159-208. https://doi.org/10.1162/qjec.122.1.159

Magnuson, Katherine A., Christopher Ruhm, and Jane Waldfogel. 2007. "The persistence of preschool effects: Do subsequent classroom experiences matter?" *Early Childhood Research Quarterly* 22(1):18-38. https://doi.org/10.1016/j.ecresq.2006.10.002

Magnuson, Katherine A., Christopher Ruhm, and Jane Waldfogel. 2007. "Does prekindergarten improve school preparation and performance?" *Economics of Education Review* 26:33-51. https://doi.org/10.1016/j.econedurev.2005.09.008

Martínez A., Claudia and Marcela Perticará. 2017. "Childcare Effects on Maternal Employment: Evidence from Chile." *Journal of Development Economics* 126: 127-137. https://doi.org/10.1016/j.jdeveco.2017.01.001

Martinez, S., Naudeau, S., Pereira, V., 2017. "Preschool and Child Development under Extreme Poverty: Evidence from a Randomized Experiment in Rural Mozambique," Policy Research Working Papers. The World Bank. https://doi.org/10.1596/1813-9450-8290

Miller, Douglas L., Na'ama Shenhav, Michel Z. Grosz, 2019. "Selection into Identification in Fixed Effects Models, with Application to Head Start." NBER Working Paper 26174. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w26174

Muschkin, Clara, Helen Ladd, and Kenneth Dodge. 2015. "Impact of North Carolina's Early Childhood Initiatives on Special Education Placements in Third Grade." *Educational Evaluation and Policy Analysis* 37(4): 478-500. https://doi.org/10.3102/0162373714559096.

Nollenberger, Natalia and Nuria Rodríguez-Planas. 2015. "Full-time universal childcare in a context of low maternal employment: Quasi-experimental evidence from Spain." *Labour Economics* 36: 124–136. https://doi.org/10.1016/j.labeco.2015.02.008

Pages, Remy, Dylan J. Lukes, Drew H. Bailey, and Greg J. Duncan. 2020. "Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts." *Educational Evaluation and Policy Analysis* 42(3): 1-22. https://doi.org/10.3102/0162373720948884

Pianta, Robert C., Karen M. La Paro, Bridget K. Hamre 2008. *Classroom Assessment Scoring System™: Manual K-3*. Paul H. Brooks Publishing.

Pihl, Ariel Marek. 2018. "Head Start and Mothers' Work: Free Child Care or Something More?" Mimeo. https://drive.google.com/file/d/1H0QbQuZg0O0Up4fk-zmShco9dqPNJdj-/view

Puma, Michael, Stephen Bell, Ronna Cook, and Camilla Heid. 2010. *Head Start Impact Study Final Report*. Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families. https://www.acf.hhs.gov/sites/default/files/opre/hs_impact_study_final.pdf

Rosinsky, Kristina L. 2014. "The Relationship Between Publicly Funded Preschool and Fourth Grade Math Test Scores: A State-Level Analysis." Master's thesis, Georgetown University.

Rossin-Slater, Maya and Mirian Wüst. 2020. "What is the Added Value of Preschool for Poor Children? Long-Term and Intergenerational Impacts and Interactions with an Infant Health Intervention." *American Economic Journal: Applied Economics* 12(3): 255-286. https://doi.org/10.1257/app.20180698

Sabol, Terri and Lindsay Chase-Lansdale. 2015. "The Influence of Low Income Children's Participation in Head Start on Their Parents' Education and Employment." *Journal of Policy Analysis and Management* 34(1): 136-161. https://doi.org/10.1002/pam.21799

Sabol, Terri J., Sandra L. Soliday Hong, Robert C. Pianta, and Margaret R. Burchinal. 2013. "Can Rating Pre-K Programs Predict Children's Learning?" *Science* 341: 845–6. https://doi.org/10.1126/science.1233517

Sall, Sean. 2014. "Maternal Labor Suppy and the Availability of Public Pre-K: Evidence from the Introduction of Prekindergarten into American Public Schools." *Economic Inquiry* 52(1): 17-34. https://doi.org/10.1111/ecin.12002

Schiman, Cuiping. 2019. "Experimental evidence of the effect of Head Start on mothers' labor supply and human capital investments." Mimeo.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, Milagros Nores. 2005. *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*. Ypsilanti, MI: High/Scope Press.

Thompson, Owen. 2018. "Head Start's Long-Run Impact: Evidence from the Program's Introduction." *Journal of Human Resources* 53(4): 1100-1139. https://doi.org/10.3368/jhr.53.4.0216-7735R1

Walters, Christopher. 2015. "Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start." *American Economic Journal: Applied Economics* 7(4): 76-102. https://doi.org/10.1257/app.20140184

Wan, Surui, Timothy N. Bond, Kevin Lang, Douglas H. Clements, Julie Sarama, and Drew H. Bailey. 2021. "Is intervention fadeout a scaling artefact?" *Economics of Education Review* 82. https://doi.org/10.1016/j.econedurev.2021.102090

Weiland, Christina, Rebecca Unterman, Anna Shapiro, Sara Staszak, Shana Rochester, and Eleanor Martin. 2019. "The Effects of Enrolling in Oversubscribed Prekindergarten Programs Through Third Grade." *Child Development*, Forthcoming. https://doi.org/10.1111/cdev.13308

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

Wikle, Jocelyn and Riley Wilson. 2020. "Access to Head Start and Maternal Labor Supply: Experimental and Quasi-experimental Evidence." Mimeo. https://economics.byu.edu/00000173-9aea-d2bd-a9f7-fbeb30dd0000/hs-laborsupply-wikle-wilson-july2020-pdf

Wong, V.C., T.D. Cook, W.S. Barnett, and K. Jung. 2008. "An Effectiveness-Based Evaluation of Five State Pre-kindergarten Programs." *Journal of Policy Analysis and Management* 27(1): 122-154. https://doi.org/10.1002/pam.20310

Yoshikawa, Hirokazu, Christina Weiland, Jeanne Brooks-Gunn, Margaret R. Burchinal, Linda M. Espinosa, William T. Gormley, Jens Ludwig, Katherine A. Magnuson, Deborah Phillips, and Martha J. Zaslow. 2013. "Investing in Our Future: The Evidence Base on Preschool Education." Report, Society for Research in Child Development, Ann Arbor, MI.

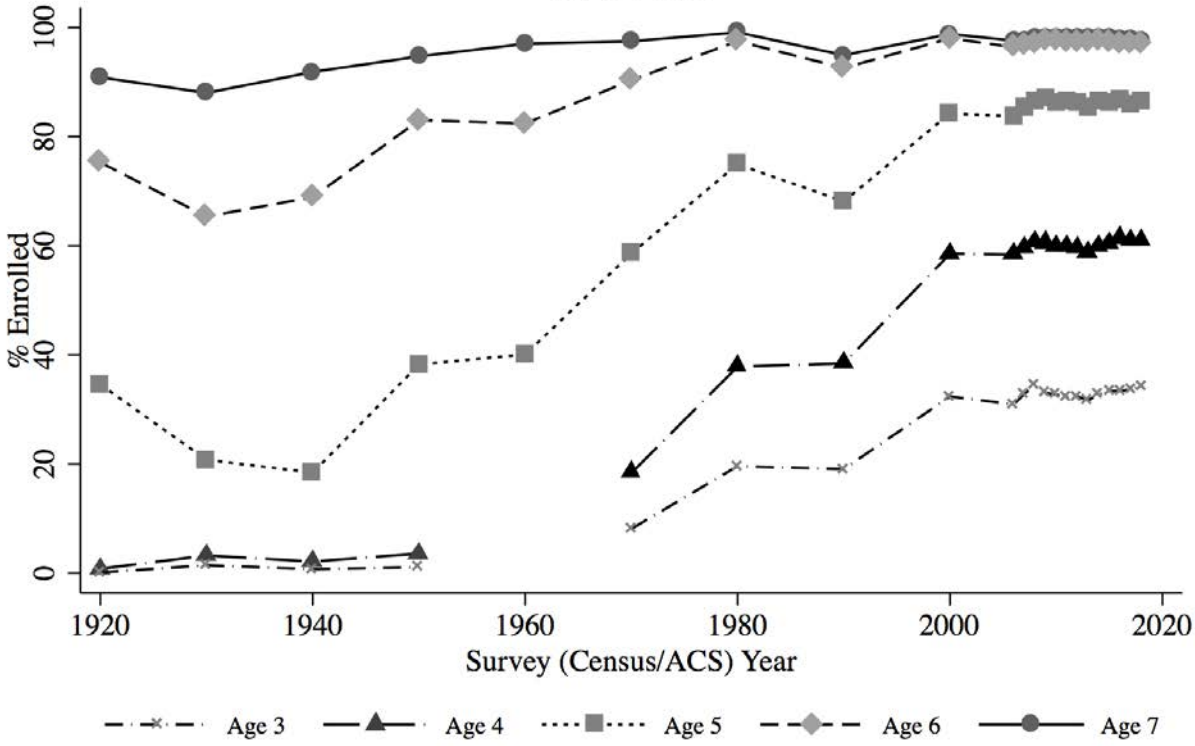
https://www.fcd-us.org/assets/2013/10/Evidence20Base20on20Preschool20Education20FINAL.pdf

Zerpa, Mariana. 2021. "Short and Medium Run Impacts of Preschool Education: Evidence from State Pre-K Programs." Mimeo. https://www.dropbox.com/s/9k81gxmnhkp8jar/Prek_latest.pdf?dl=0

Zhai, Fuhua, Jeanne Brooks-Gunn, and Jane Waldfogel. 2014. "Head Start's Impact Is Contingent on Alternative Type of Care in Comparison Group." *Developmental Psychology* 50(12): 2572-2586. http://dx.doi.org/10.1037/a0038205

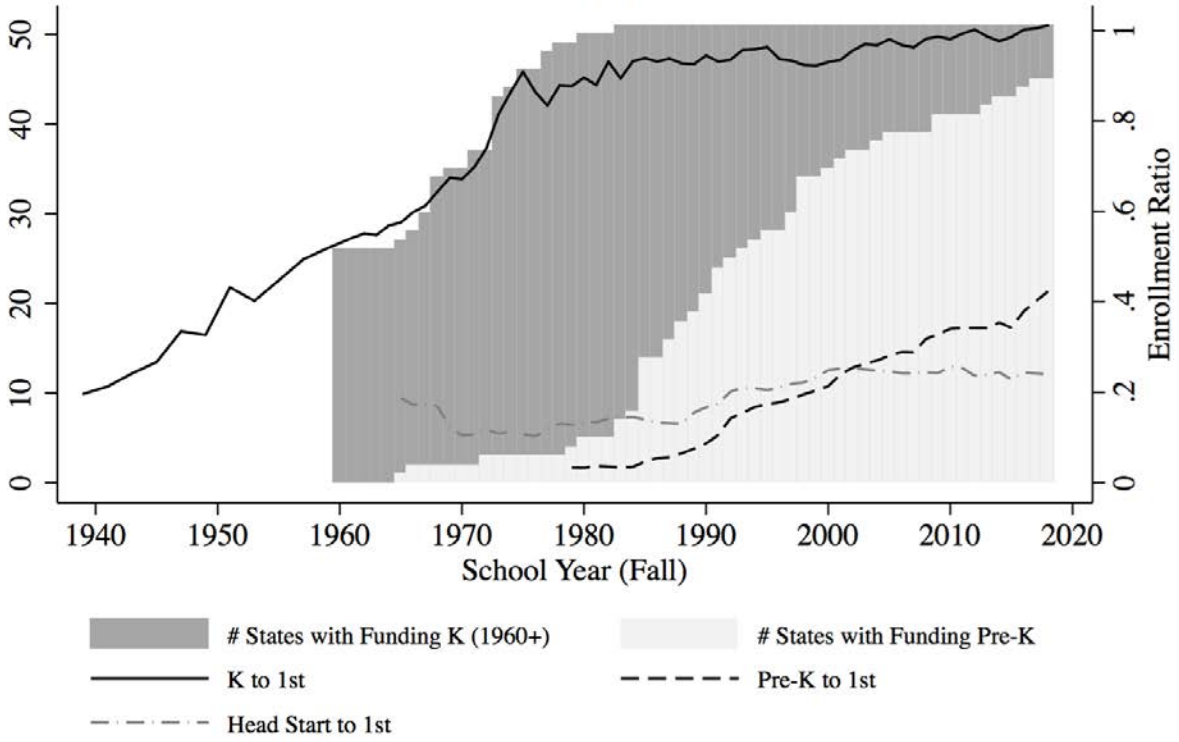
Zigler, Edward F. and Susan Muenchow 1992. *Head Start: The Inside Story of America's Most Successful Educational Experiment.* New York: Basic Books.

Figure 1. U.S. School Enrollment Rates by Age and Year: 1920-2018



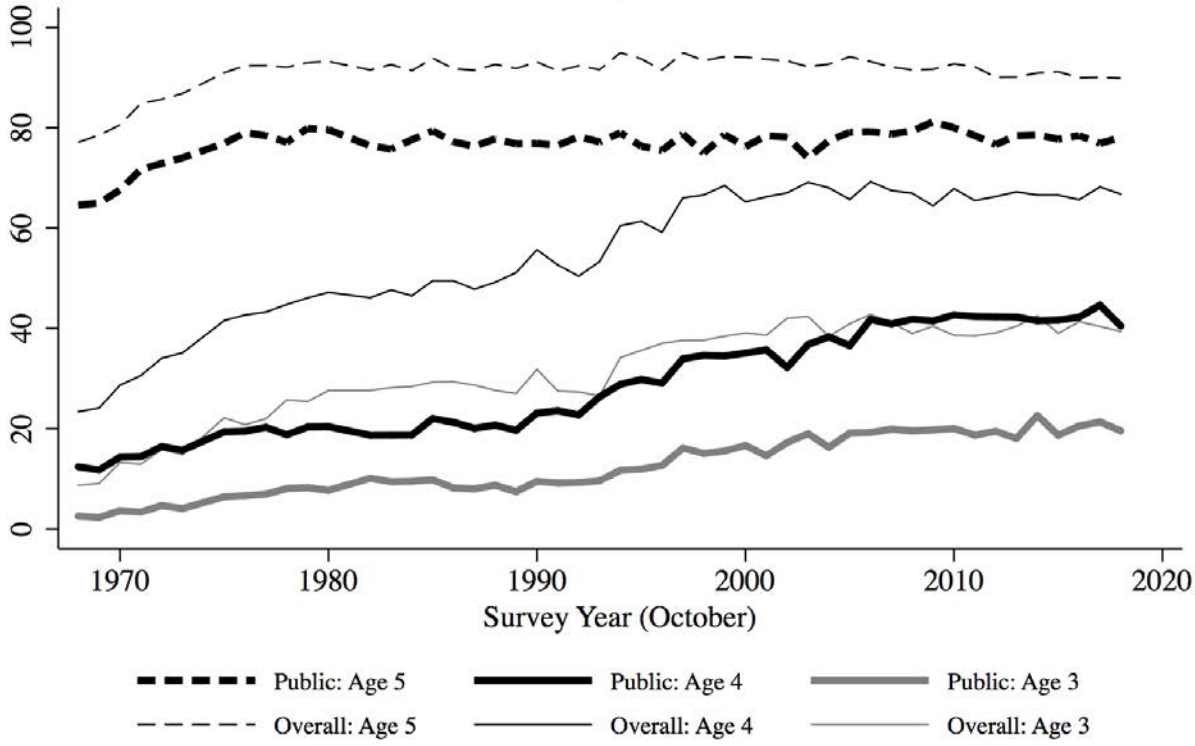
Sources: Public use microdata samples from the Decennial Census (1920-2000) and American Community Survey (2006-18) (Ruggles et al., 2020).
Notes: Author's calculations limiting sample to children residing in the 50 states or Washington, D.C. and for whom age and school enrollment are not allocated and weighting by person weights (sample line weights in 1950). School enrollment incorporates both public and private education and is not reported for children under the age of five in 1960.

Figure 2. U.S. Public Early Education Enrollment Ratios by Year, 1939-2018



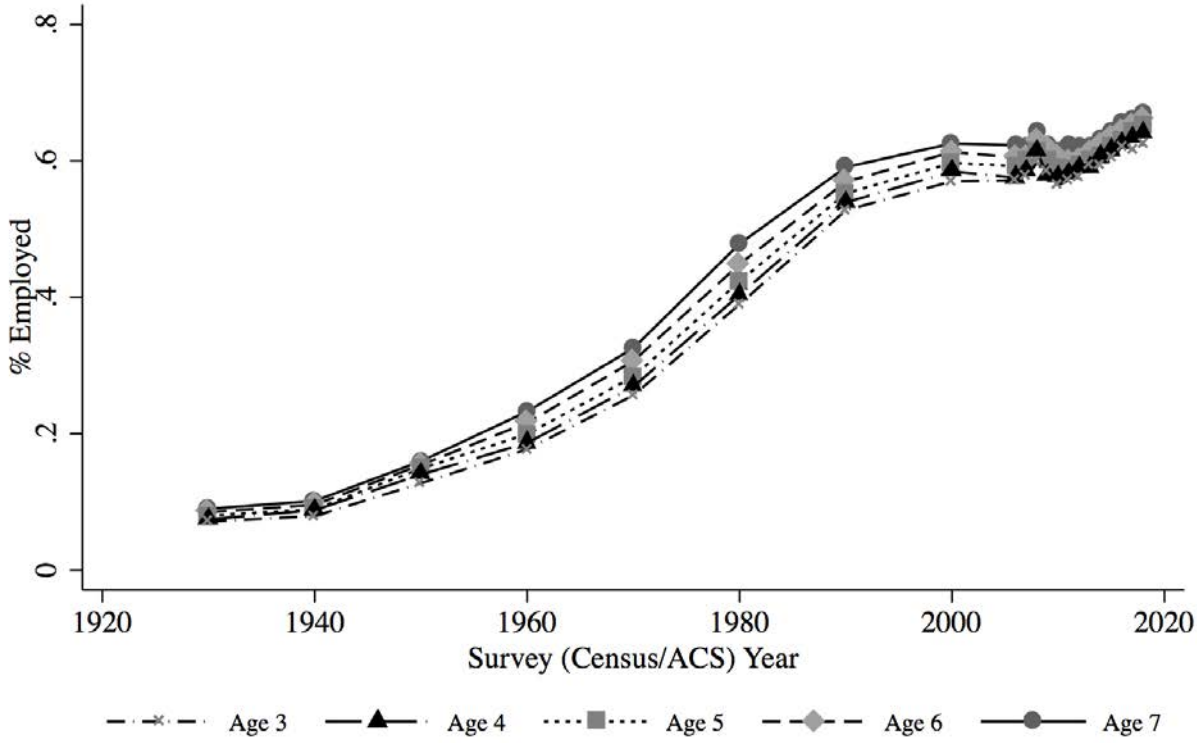
Sources: For public school enrollment: *Biennial Survey of Education* (1939-57), *Statistics of State School Systems* (1958-63), *Statistics of Public Elementary and Secondary Day Schools* (1964-67), *Fall Statistics of Public Schools* (1968-76), *Statistics of Public Elementary and Secondary Schools* (1977-80), *Public School Enrollment, United States* (1981-82), *Common Core of Data: State Nonfiscal Survey* (ICPSR Study No. 6947), Common Core of Data "State Non-fiscal Public Elementary/Secondary Survey" (1986-2018) (http://nces.ed.gov/ccd/elsi/). For Head Start enrollment: "Head Start Federal Funding and Funded Enrollment History" (https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/head-start-federal-funding-funded-enrollment-history-eng.pdf). For kindergarten funding: Cascio (2009a, 2009b). For pre-kindergarten funding: National Institute for Early Education Research.
Notes: Author's calculations. Enrollment figures for pre-K, kindergarten, and first grade are in public schools. Kindergarten funding dates are left censored at 1960.

Figure 3. U.S. School Enrollment Rates by Age and Survey Year: October CPS, 1981-2018



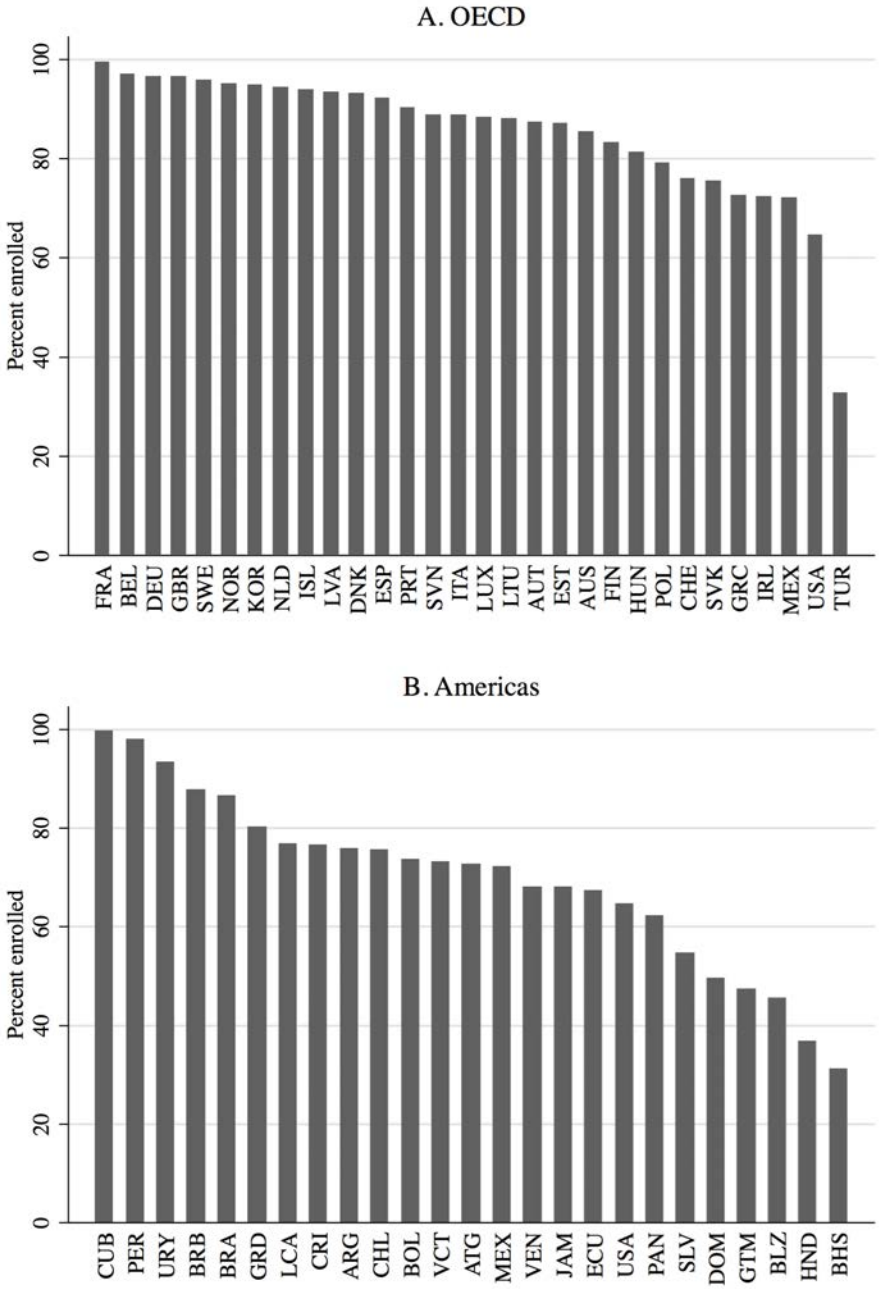
Sources: Public use microdata samples from the October Current Population Survey School Enrollment Supplement (Flood et al., 2020).
Notes: Author's calculations limiting sample to children residing in the 50 states or Washington, D.C. and for whom age and school enrollment are not allocated and weighting by final sampling weights.

Figure 4. U.S. Maternal Employment Rates by Age of Child and Year: 1930-2018



Sources: Public use microdata samples from the Decennial Census (1930-2000) and American Community Survey (2006-18) (Ruggles et al., 2020).
Notes: Author's calculations limiting sample to children residing in the 50 states or Washington, D.C. and for whom child age and school enrollment and maternal employment are not allocated and weighting by (child) person weights (sample line weights in 1950).

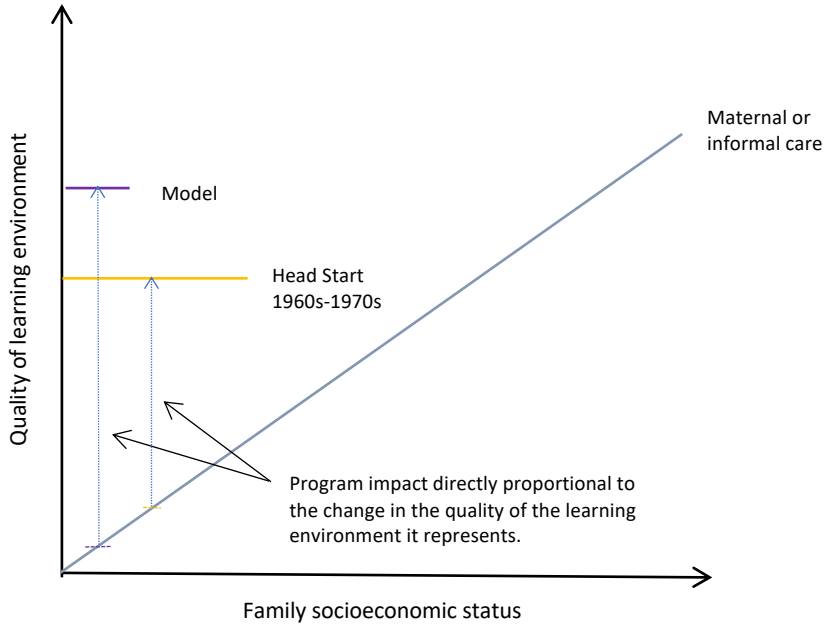
Figure 5. Enrollment Rates in Early Education Across the World



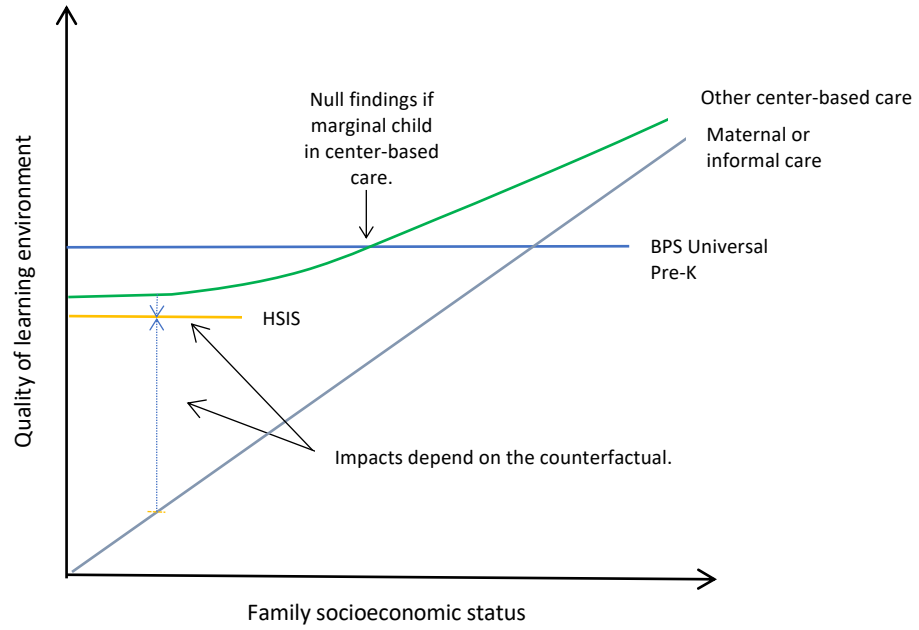
Source: UNESCO Institute for Statistics (http://data.uis.unesco.org).
Notes: Figures shown are net enrollment rates in pre-primary education (NER_02_CP). The enrollment rate for Great Britain (GBR) corresponds to 2015. Some countries under a given classification have no data reported. For example, Panel A is missing data on Canada, Colombia, Czech Republic, Israel, and Japan.

Figure 6. Conceptual Framework for ECE Impact Evaluation

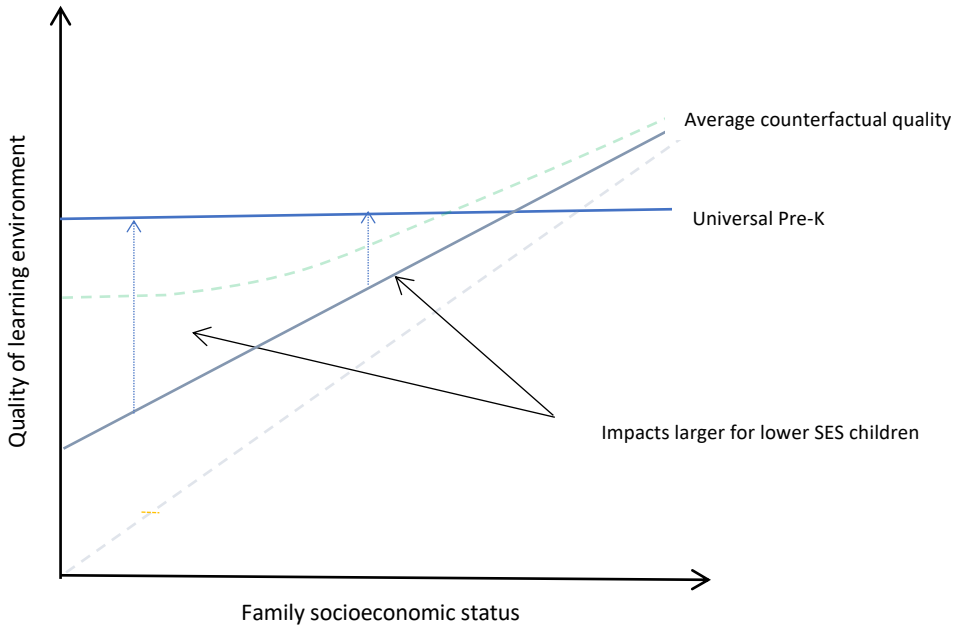
A. Counterfactual is Maternal or Informal Care: Targeted Programs



B. Counterfactual with Other Center-Based Care: Targeted Programs



C. Counterfactual with Other Center-Based Care: Universal Pre-K



Source: These figures are adapted from figures presented in Cascio and Schanzenbach (2014).

Table 1. Gaps in Overall School Enrollment across Groups Defined by Race, Region, and Family Background: by Age and Year

	Year							
	1940	1960	1970	1980	1990	2000	2010	2018
				A. Region (South - non-South)				
Age 5	-20.4	-32.0	-22.6	-4.3	-2.0	-0.3	1.6	0.5
Age 4	-0.8		-2.7	-3.3	-3.3	-0.6	-2.2	-2.9
Age 3	-0.2		0.6	2.4	2.2	3.7	-1.0	-3.2
				B. Race (Nonwhite - White)				
Age 5	-7.7	-2.4	-4.4	3.8	2.0	1.6	2.1	1.4
Age 4	0.0		2.9	8.1	-0.7	-1.0	0.3	2.1
Age 3	0.0		1.1	4.8	-1.5	1.4	1.7	0.4
				C. Parental Education (Below Median - Above Median)				
Age 5	-7.8	-12.6	-17.6	-11.6	-11.1	-6.9	-6.2	-5.2
Age 4	-0.4		-18.3	-22.8	-20.2	-17.9	-18.2	-17.9
Age 3	0.2		-9.2	-17.5	-15.2	-16.5	-21.1	-22.0
				D. Maternal Employment (Mother Not Employed - Mother Employed)				
Age 5	-1.6	-0.6	-1.5	-6.2	-4.5	-4.3	-4.4	-4.3
Age 4	-1.3		-5.0	-12.1	-7.8	-9.6	-8.5	-10.3
Age 3	-0.1		-6.0	-12.3	-7.4	-10.9	-11.9	-12.7

Sources: Public use microdata samples from the Decennial Census (1920-2000) and American Community Survey (2006-18) (Ruggles et al., 2020).

Table 2. Gaps in Public School Enrollment across Groups Defined by Race, Region, and Family Background: by Age and Year

	Year						
	1960	1970	1980	1990	2000	2010	2018
			A. Region (South - non-South)				
Age 5	-34.3	-30.8	-5.7	0.0	0.9	4.3	2.3
Age 4		-3.1	-3.0	-1.6	-0.2	1.1	-0.3
Age 3		0.1	-0.6	0.3	0.4	-1.5	-2.4
			B. Race (Nonwhite - White)				
Age 5	-1.5	3.7	10.9	12.7	14.2	10.7	8.5
Age 4		8.4	14.8	10.6	12.6	9.4	10.0
Age 3		2.7	6.2	3.9	6.7	6.5	5.8
			C. Parental Education (Below Median - Above Median)				
Age 5	-8.4	-2.9	6.5	9.0	17.2	16.3	15.8
Age 4		0.9	2.6	3.8	11.0	13.3	12.5
Age 3		0.0	-0.6	-0.7	2.6	2.9	3.0
			D. Maternal Employment (Mother Not Employed - Mother Employed)				
Age 5	0.4	1.3	0.4	1.9	0.0	-0.9	-1.9
Age 4		-1.5	-1.9	-0.1	-1.9	1.0	-2.1
Age 3		-1.9	-2.1	-1.7	-3.2	-2.5	-3.5

Sources: Public use microdata samples from the Decennial Census (1920-2000) and American Community Survey (2006-18) (Ruggles et al., 2020).