THE UNIVERSITY OF CHICAGO

THREE ESSAYS ON THE ECONOMICS OF EARLY CHILDHOOD EDUCATION PROGRAMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECONOMICS

BY

ANDRES PABLO CRISTOBAL HOJMAN CRUZ

CHICAGO, ILLINOIS

AUGUST 2016

A Melé, Keno y Doren, que forjaron mi educación desde mi más temprana infancia, y a mi Yumi, que ha sido una bella compañía durante todos estos años.

"When you are studying any matter, or considering any philosophy, ask yourself only what are the facts and what is the truth that the facts bear out. Never let yourself be diverted either by what you wish to believe, or by what you think would have beneficent social effects if it were believed. But look only, and solely, at what are the facts."

- Bertrand Russell

Contents

List of Figures

List of Tables

Chapter 1

New Evidence for an Old Puzzle: The Fadeout of Early Childhood Education Impacts on Cognitive Skills

1.1 Introduction

Early childhood education programs are key policy tools that can have high economic returns and equalize opportunities in society.[1] Although many of these programs present strong initial impacts on cognitive test scores, several evaluations show that these effects dissapear or strongly diminish shortly after the end of the programs.[2] This phenomenon is known as "fadeout", and it is the main criticism of early childhood education programs in policy debates.[3] Despite its prevalence, its causes remain unknown and largely unstudied. In this paper I describe the timing of the fadeout and identify its possible causes.

In the first part of the paper, I describe the main trends in the data and the timing of the fadeout. I use two *preschool* programs, the Perry Preschool Program (Perry) and the Early Training Project (ETP).[4] In these datasets, participation is exogenous because the programs were randomly assigned, and there are no treatment substitutes. The main trends in the data provide some insights about the factors underlying the fadeout phenomenon: The raw scores show that the level of the children's skills are increasing with age across all the periods covered in the data. The standardized scores show that the skills of children in the control group have substantial increases at school entry compared to the average US children, even controlling for age.[5]

[1]This paper focuses on high-quality education/stimulation programs targeted at disadvantaged children ages 0-5. Positive long-term impacts of these programs are discussed in Blau and Currie (2006a), Heckman et al. (2010a), Campbell et al. (2014), and Elango et al. (2015). Two comprehensive cost-benefit analyses are Heckman et al. (2010b) and García et al. (2016)

[2]Appendix A shows further evidence of this phenomenon as compared across several early childhood education programs. Throughout the paper I use the term "cognitive test scores". Most of the tests that I use aim to measure what is commonly known as Intelligence Quotient (IQ). While there might be more long-term impacts on achievement test scores, we know little about the fadeout of achievement tests because they are generally not available until children have entered school. Heckman and Kautz (2012a) discuss the difference between IQ and achievement tests.

[3]Some representative articles are Dalmia and Snell (2008), Guernsey and Bornfreund (2013), Alison (2014), Armor (2014), and Kirp (2015). These critiques have been present since the first results of school-age follow-ups of early childhood education programs were published (Project Head Start, 1969; Lazar et al., 1982).

[4]I define Preschool programs as education/stimulation programs attended by children during one or two years before they enter school (ages 3–6 in these data).

[5]There is variation in age of school entry in these datasets: When Perry was implemented, Kindergarten was part of public schools on its state. When ETP was implemented, public education started on first grade

I study the period-by-period change in the gap between the treatment and the control groups to identify the impacts of each period's events. I find patterns of impacts that were not previously discussed in the literature: (i) around 63% of the maximum gap between the groups is created within the first two months of the programs; (ii) a second year of exposure to the programs does not generate additional impacts; and (iii) after the programs end, around 83–99% of the fadeout happens in the first year of school. There are two major changes for the children in that year: they exit preschool and they enter school. In principle, any of these two factors, or both, could be causing the fadeout.

In the second part of the paper, I separate the program-exit effect from the subsequent-schooling effect. We could disentangle them if we had a second experiment after the programs end, whith children being randomized into attending schools or staying at home. Given that schools are generally universal, I propose an alternative approach by studying fadeout of *toddler care* (ages 1–3) impacts at preschool entry, instead of fadeout of preschool impacts at school entry.[6]

I use data from a program that provided toddler care at ages 1–3, the Infant Health and Development Program (IHDP). The data has the unique feature that after a first period when the randomized program is in place, there is a second period when families can choose whether to enroll children in preschool or not. This provides the necessary group that attends a program and then stays at home in a second period, so their test scores are affected by program exit, but not by subsequent schooling.

Given that assignment in the second period is not randomized, I take two steps to avoid endogeneity. First, I only estimate *within* children attending preschool or *within* children not attending. Second, within each of these two groups, the ones who were randomized into IHDP could in principle be different from those who were randomized out of it. I document that in practice the randomization did not significantly affect preschool take-up in the data, so it is possible to assume that the composition of children attending preschool is not altered

on its state.

[6]I define *toddler care* as education/stimulation programs for children ages 1–3.

by randomization.

Then, I am able to identify the pure program-exit effect by studying the impact of the randomization into IHDP in children that did not attend preschool. I find that within this group there is no fadeout in the second period. The impacts do not diminish until after children enter schools. On the other hand, there is a strong fadeout for the IHDP treatment group that attend preschool: most of the advantage acquired up to age 3 is gone by age 5.

If the underlying mechanisms for fadeout at early ages can be extrapolated to school entry, then the fadeout phenomenon is caused by the positive school-entry effect in control children, not the negative preschool-exit effect in treated children. Thus, opposite to a widely held belief, my findings suggest that depreciation is not empirically relevant for fadeout. This finding contributes to understand the old puzzle of the disappearance of the impacts on cognitive skills.

The rest of this paper is structured as follows: Section 1.2 discusses some hypotheses about fadeout in the literature. Section 1.3 discusses conditions to interpret my findings in terms of skills. Section 1.4 presents the three datasets I use and the main trajectories of cognitive skills in the data. Section 1.5 explains how a skill formation framework allows me to identify the timing of the impacts. Section 1.6 shows the results from the estimations. Section 1.7 concludes.

1.2 Literature Discussing Fadeout and Hypotheses

1.2.1 Descriptions of Fadeout

The existence of the fadeout phenomenon is well known in the literature. Currie (2001a), Barnett (2011) Duncan and Magnuson (2013a), Elango et al. (2015) and Bailey et al. (2015) are a few of the many papers documenting it. There are a few remarkable examples of programs that present gains on IQ using relatively long-term measurements. However, even for these exceptions, the impacts during the program are much stronger than the long-

term impacts (Campbell et al., 2002; Duncan and Sojourner, 2013a). Later in the paper I show that the raw test scores have increasing trends across all years. This is also relatively well known. Based on that, some studies prefer to describe the fadeout as "catch-up" of the control group (Duncan and Magnuson, 2013a; Yoshikawa et al., 2013a). Four papers, Camilli et al. (2010), Leak et al. (2010), Various (2014) and Bailey et al. (2015), use meta-analyses to describe how program impacts decrease after the end of the program. They respectively estimate the decline to be 10, 20, 33 and 56% of the total impact of the program per year (the last two explicitly calculate the decline after the first year). In a recent paper, Bailey et al. (2015) discuss the types of skills that could generate long-term impacts in children, from the point of view of three different theoretical approaches.

Even the largest of the estimates about the decline of impacts in the literature is far from my estimates of 83–99% of the fadeout closing in the first year of program.[7] For the case of cognitive skills, I provide evidence contrary to what Bailey et al. (2015) call the "sustained environments perspective", as in my data higher quality environments for both the control and the treatment groups imply stronger fadeout. Magnuson et al. (2007) is fully consistent with my findings that fadeout is occurs when children are exposed to subsequent schooling.

The most closely related paper to this study in the education literature is Magnuson et al. (2007). They use nationally representative data and present the surprising finding that fadeout is *stronger* in higher-quality schools. The scenario I study, the absence of subsequent education after attending a program, might be considered a more extreme version of the low-quality schools they study. I improve on their work by analyzing the timing of the fadeout, by testing explicit hypotheses, and by correcting for selection on unobservables into preschool and into subsequent experiences. The latter is especially important, as it is hard to determine if the impacts they find are distorted by the composition of the groups.

[7]This difference might be due to the presence of treatment substitutes in the programs considered in the meta-analyses. This makes the bulk of the impacts close while the program is still running, making the impacts that remain after the second year of school a larger fraction of the impacts at the end of the programs.

1.2.2 Hypotheses About the Causes of Fadeout

Almost all surveys on early childhood education programs mention some hypotheses about the causes of fadeout. However, there are surprisingly few studies formally testing those hypotheses. Thus, the exact meaning that different authors give to them is not always obvious.

- *Hypothesis Related to the Loss of Previous Gains After Program Exit.* The most common theory for fadeout is that gains from preschool are lost because children exit the programs and they do not receive substantial positive investments after that.[8] Most often, the authors hypothesize that gains from the program could be maintained if high-quality investments were in place (Currie and Thomas, 1995a; Barnett, 1995, 2011; Duncan and Magnuson, 2013a; Yoshikawa et al., 2013a; Bailey et al., 2015). They are based on the assumption that the schools that children participating in the experimental programs attended were very low quality. Two interpretations of this theory could have different observable implications: The first is that there will be depreciation of the skills gained from the program. If this is true, there would be a negative impact on the skills of treated children after program exit (fadeout of *skills* rather than just fadeout of *gains*). This negative impact could not be observable in the data trends, if it is masked by positive age effects. The second interpretation is that even in absence of high quality formal education, control children would learn what treated children learned earlier in early childhood education programs. In that case, there will be no negative trends in the data.[9] As an example of this group of hypotheses, in Appendix

[8] In this paper, I operationalize substantial positive investments as participation in a formal early childhood education program. Given the high impacts of the preschool programs I analyze, we know that, at least for cognitive test scores, and in these disadvantaged populations, these programs are more effective than family/informal care for fostering cognitive skills.

[9] Note that the Depreciation hypothesis is valid if *all* children attend low-quality schools. An alternative explanation,that has been suggested for the fadeout observed in Head Start, is that children *that participate on it* attend lower quality schools than comparison children (Lee and Loeb, 1995; Currie and Thomas, 2000). While that explanation can account for part of the fadeout in studies using observational data, I document that in the experimental programs I consider treated and control children attend similar schools, and the data still show strong fadeout.

B I present a formal model of depreciation, in which the gains from early childhood education programs depreciate faster than other cognitive skills.

- *Hypotheses Related to Differential Gains from Subsequent Schooling.* Several hypotheses given in the literature state that control group children gain more from subsequent schooling experiences than treated children. There are at least two reasons why this can be true. One is that teachers make compensatory efforts to help the least prepared children catch-up. If this is true, as treatment children would be more school-ready than control children, the former would get less attention from their teachers (Barnett, 2011; Duncan and Magnuson, 2013a).[10] A second reason is that there might be diminishing returns to education on the production of test scores: if a few easily trainable concepts are necessary to answer a test, only the first exposure to formal education will make a large impact on test scores. If that is true, treatment children will obtain that boost during their first year in preschool, and control group children will obtain it at school entry, and fadeout will be observed in the data. As an example, in Appendix B, I present a formal model of differential gains from subsequent education.

- *The Statistical Artifact Hypothesis.* In one of the few papers that directly attempts to test a hypothesis for fadeout, Cascio and Staiger (2012) discuss the possibility that part of the observed fadeout on test scores is due to using measurements that are age-standardized, which could imply presenting estimates in terms of standard deviations that increase period by period. In this case, a permanent impact could look smaller across time because the denominator of the age-standardization is increasing in each period. The authors assess this possibility and find that it does play a role in fadeout. Although I use standardized scores in the main estimates, Appendix J shows that my findings are robust to using other transformations of the measurements that are unaffected by this statistical artifact.

[10] Two observational papers (Engel et al., 2013; Claessens et al., 2013) find that Kindergarten teachers spend most of their time teaching the most basic concepts, which are only helpful for the least prepared students.

In the paper, I test the empirical support of the subsequent-schooling hypotheses compared to the support of the program-exit hypotheses, finding that the latter have little support in the data. In Section 1.7, I discuss what are the most plausible mechanisms behind differential gains from subsequent schooling.

1.3 Measurement of Skills

In Section 1.2, I mention several papers that give evidence on the existence of fadeout on test scores. However, the most important policy question is whether there is fadeout on the impacts on skills. While scoring better in a test can have little consequence, having a higher level of cognitive skills can substantially improve an individual's life prospects. Thus, in this paper, I take three steps toward using skills, not just measurements, as the variables of interest: (i) I account for measurement error; (ii) I use statistics that are invariant to the arbitrary scale of test scores, and (iii) I check the robustness of my estimates using several transformations of the test scores.

Throughout this paper I generally assume that tests have quantitative information. This is a strong assumption, which some economists have critisized (Cunha et al., 2010a; Bond and Lang, 2013; Jacob and Rothstein, 2016). Given that my research question involves quantifying magnitudes of skills, I need to take this assumption for my main estimates.[11] In Appendix J, I test the robustness of my main estimates using several transformations of the test scores, and different types of tests, finding that the qualitative patterns hold generally across almost all of the transformations.

[11] However, in Appendix F, using only ordinal properties of the test scores, I can still identify some patterns on skills. In particular, (i) in terms of the levels, I show an increase in the skills for both groups across all periods; and (ii) in terms of the gaps between the treated and the control groups, I show that the data is consistent with fadeout on skills. The main findings are that at baseline, there is no stochastic dominance between the treatment and the control group. Then, during the program, there is stochastic dominance of the treatment group. Finally, in later periods there is no stochastic dominance. Moreover, at later ages I cannot reject the hypothesis that the distributions are identical, while it is rejected during the program. I also show that the distribution of the test scores at each age dominates the distribution of skills in the previous age for the control and for the treatment groups. Those patterns match with the dynamics obtained using quantitative properties for the tests, as shown in Section 1.4.

1.3.1 Measurement Model

I define my object of interest in this paper as the (possibly narrow) type of cognitive skills that test scores measure. I assume that the tests are dedicated measurements of skills, so all tests that I use measure a single scalar-valued skill. Let a measurement of cognitive skills for individual i at time t be M_{it}^m. The type of test, m, refers to each of the different cognitive tests in the data.[12]

I assume a linear measurement model for the relationship between test scores and skills. This is the standard model in the economics literature, and it arises naturally when using quantitative properties of test scores.[13]

$$M_{it}^m = a^m + b^m \theta_{it} + \varepsilon_{it}^m, \tag{1.2}$$

where ε_{it}^m is the measurement error. I also assume that the measurement errors are independent from the skills, and have a mean of zero: $\theta_{it} \perp \varepsilon_{it}^m$, $E[\varepsilon_t^m] = 0$.

In this paper I combine different types of tests to maximize statistical power and to observe changes across finer periods of time. Thus, I use transformations of the measurements for which the assumption of identical slope parameters is plausible (for example, it is not reasonable to assume that the b parameters are identical for Raw Scores of different types of tests, because in practice they have different scoring scales).[14] In particular, for my

[12] In section 1.4, I discuss the available types of tests. Two examples of types of tests are the Stanford-Binet Test and Peabody Picture Vocabulary Test (PPVT)

[13] This measurement model is common in the economics literature when papers attempt to draw conclusions about skills (Todd and Wolpin, 2003, 2007; Cunha and Heckman, 2008a). Let a generic non-stochastic version of a measurement be $M_{it}^m = g(\theta_{it})$. The literature on measurement theory organizes and rigorously describes several types of scales with different properties, and the weakest assumption necessary for measurements to have quantitative properties is the preservation of intervals (Narens, 1981; Krantz et al., 1971; Iverson and Falmagne, 1985; Luce and Narens, 2008). Let a set of four measurements in any subject-time combination be indexed by $\{1, 2, 3, 4\}$. I define that a non-stochastic measurement M_{it}^m preserves intervals if:

$$M_1^m(\theta_1) - M_2^m(\theta_2) \gtreqless M_3^m(\theta_3) - M_4^m(\theta_4) \iff \theta_1 - \theta_2 \gtreqless \theta_3 - \theta_4. \tag{1.1}$$

If the measurement preserves intervals, it is possible to show that the relationship between the measurement and the test will be affine, so $M_{it}^m = a^m + b^m \theta_{it}$ with $b^m > 0$. Given that, the linear measurement model in Equation (1.2) seems like a natural assumption.

[14] I define all transformations mentioned in the paper formally in Appendix G

main estimates I use: (i) Raw Scores of a single type;[15] (ii) Standardized Scores, which are expressed in standard deviations of the national raw scores at each age; and (iii) Mental Age Scores, which are an estimate of the average age for children that have a given score in the test. For all of these tests, I assume that their b^m parameter is the same. This is a strong assumption for (ii) and (iii), so I check how my estimates change when I only use test scores of the same type in Appendix J. Combining multiple tests generally make trends to look less smooth, but all the main insights are robust.

1.3.2 Parameters of Interest in this Paper

Let the Randomization into the programs I use in this paper be R_i. A child randomized into a program has $R = 1$, and $R = 0$ otherwise. An important assumption that I require in this paper is that $E\left[\varepsilon_t^m | R\right] = 0$. This assumption is not trivial and could not be satisfied in practice. It can be interpreted as assuming that tests are not artificially inflated by the treatment, so that gaps in the tests reflect gaps on skills.[16] If this assumption is true, the measurement model allows me to identify the change in gaps on skills up to a scalar using the change in gaps on measurements. Most estimates in this paper are based on these statistics. Let ΔV be defined, for any variable V, in the following way: $\Delta V \equiv E[V|R = 1] - E[V|R = 0]$. Then, we can construct:

$$\Delta M_t^m - \Delta M_{t-1}^m = b^m \left(\Delta\theta_t - \Delta\theta_{t-1}\right) \tag{1.3}$$

From model (1.2), we can conclude that it will only be possible to identify the moments

[15] In practice, when tests are of the same type, the available questions are generally the same for each year. The children do not answer exactly the same each year because, while giving the tests, the examiners skip the questions that are considered too easy for the capacity of the children and stop when they are too hard. The type of test, m, is important because it is possible to assume that tests of the same type will have the same b^m parameter.

[16] The plausibility of this assumption depends on the breadth of the definition of the skills we use. For example, if the programs makes treated children better at tests because they teach them very simple skills that are useful for taking tests as names of common objects, and the definition of skills includes those skills, the assumption will be satisfied. If those skills are excluded, the assumption could not be satisfied. For this paper I have to assume a broad interpretation.

of θ up to a scalar. That is natural, as it should not be possible to identify the scale of the skill. Given that problem, I report my main findings in terms of ratios using other statistics. In particular, I report the changes in the gap on skills across two periods relative to the total fadeout in the data. For example, under the linear measurement model it is possible to claim that 83–99% of the fadeout happens at the first year of school. If the period of the maximum gap in measurements is P and the period of the minimum gap is T, then I define the fadeout as $\Delta\theta_P - \Delta\theta_T$.[17] Then, the change in the gap between periods t and $t-1$ can be expressed as:

$$\frac{\Delta M_t^m - \Delta M_{t-1}^m}{\Delta M_P^m - \Delta M_T^m} = \frac{\Delta\theta_t - \Delta\theta_{t-1}}{\Delta\theta_P - \Delta\theta_T}. \tag{1.4}$$

We can construct the left-hand side. In theory, this statistic will be invariant to the arbitrary choice of the tests because it does not depend on the test-specific parameters a^m and b^m. In practice, the estimated magnitudes will have sample variation and will change from test to test.

1.4 Data

The data in this paper comes from three early childhood education programs: (i) the Early Training Project (ETP); (ii) the Perry Preschool Project (Perry); and (iii) the Infant Health and Development Program (IHDP). The first two datasets are very similar, and I use them separately from IHDP. The ETP and Perry datasets share some characteristics that make them ideal for the study of the patterns of impacts: Both were implemented using randomized controlled trials (RCTs) with minimal compromises to randomization. Additionally, there were no treatment substitutes for the control children. The treatment for both programs was dispensed in two groups of cohorts, with one group receiving two years of preschool and one group receiving one. Many rounds of cognitive tests were collected from the subjects before, during, and after the actual programs.

[17] In practice, I also remove any baseline differences from the maximum gap in measurements.

While the data in ETP and Perry are very rich, they do not have variation in exposure to formal education after children leave the program: all children started attending school immediately after the programs ended. That is the type of variation that is needed to separate the Differential Gains Model from the Depreciation Model. The IHDP data gives access to this type of variation. Table 1.1 illustrates the timing of the program for the different experimental groups and the test applications for all three programs.

1.4.1 Early Training Project

The Early Training Project (ETP) was implemented in Abbotsfield, Tennessee from 1962 through 1966. The program was intended for children aged four to five years old prior to entering public school (Gray et al., 1982a). The targeted children were all African American, and they were considered disadvantaged as defined by family income, housing characteristics, and maternal characteristics including education (at most eighth grade) and occupation (unskilled or semi-skilled, or unemployed) (Klaus and Gray, 1968).

Due to segregation, all African-American children in the community, including all ETP participants, attended the same elementary school. In Tennessee at the time, public school began with first grade (Gray et al., 1982a; Cascio, 2009). All the teachers were African American, with education and experience comparable to those of the white teachers in other schools (Gray and Klaus, 1970).

There were sixty-one children randomized into one control and two treatment groups. The two treatment groups received different durations of the program: one of the groups received three summers of center-based care and two winters of home visits starting at age four. The other treatment group only received two summers of center-based care and one winter of home visits starting at age five (Klaus and Gray, 1968).

The data contain four different IQ measures administered with varying degrees of frequency.[18] Aggregating all the IQ measures, there is at least one IQ measure before and after

[18]The tests are ITPA, PPVT, SB, and WISC.

each of the three summers of treatment, as well as before and after school entrance with additional follow-ups at ages 7, 8, 10, and 16.[19]

1.4.2 Perry Preschool Project

The Perry Preschool Project (Perry) was implemented in Ypsilanti, Michigan, from 1962 through 1967. To be eligible, children needed to be African-American, have an IQ ranging from 70 to 85 (which is low compared to the national mean of 100), and come from a disadvantaged family as defined by parental employment, income, education, and housing characteristics (Weikart et al., 1967).

All of the Perry participants attended Perry Elementary School. In Michigan at the time, public school began with kindergarten. Most of teachers at this school were also African-American and had at least a bachelor's degree (Berrueta-Clement et al., 1984).

There were 123 children who participated in Perry; of these children, 58 were randomly assigned to the treatment group and 65 to the control group (Weikart et al., 1967). These children were assigned to five cohorts based on birth date. The treatment included a 2.5-hour preschool session on weekdays during the school year and weekly home visits from their teachers lasting 1.5 hours. The program lasted for two years starting at age three except for children in the first wave, who were all four years of age upon entry and who only received one year of treatment (Weikart et al., 1978).

Data on the subjects were collected annually while they were ages 3 through 11, and again at ages 14, 15, 19, 27, and 40. Various IQ and achievement measures were used to assess the cognitive abilities of subjects over time. Five IQ tests were administered with varying degrees of frequency to test subjects at various ages.[20] For a full description of the cognitive measures in the data, see Appendix C.

[19]For a full discussion of the cognitive measures present in the data, see Appendix C.
[20]The tests are ITPA, Leiter, PPVT, SB, and WISC.

1.4.3 The Infant Health and Development Program

The Infant Health and Development Program (IHDP) was a large-scale randomized controlled trial aiming to study the development of premature, low birth-weight children. It took place in the medical facilities of eight major university campuses during the mid-to-late 1980s. The children born in those clinical sites were eligible to be participate if they weighted 2.5 kilograms or less, had a maximum gestational age of 37 weeks, and lived within 45 minutes of the sites.

The 985 participants were stratified by clinical site and birthweight group ($\leq$2 kg or 2-2.5 kg). One third was randomly assigned into the treatment group and two thirds into the control group.

The treated group received home visits immediately after birth (weekly in first year, biweekly from ages one to three), which aimed to implement a curriculum for the child, as well as to teach parents problem-solving and child-rearing skills. After the first year, the treated children were required to enroll in center-based childcare. This component was freely provided by IHDP for 5 days a week, between 4-9 hours per day, for 24 months. The childcare centers had teacher:child ratios from 1:3-1:4, and continued the curricula from the home visits. From ages one to three, the treated group also had access to bimonthly parent meetings for parents to discuss concerns and provide support to each other (Gross et al., 1990; Brooks-Gunn et al., 1992; Martin et al., 2008).

1.4.4 Cognitive Scores Trajectories in the Data

1.4.4.1 Raw Scores Show Skills Generally Increase Across Time

It is possible to identify the sign of the trends in skills from the trends in scores, if repeated measurements of the same tests are available. From Equation 1.2, we have:

$$E\left[M_{it}^{m}\right] - E\left[M_{it-1}^{m}\right] = b^{m}\left(E\left[\theta_{it}\right] - E\left[\theta_{it-1}\right]\right) \tag{1.5}$$

Table 1.1: Timing of Formal Education, Perry, ETP and IHDP

Perry	Age (months)			
	0–36	37–48	49–60	61–onwards
Treated 2-year Program				
Treated 1-year Program				
Control				
Tests	36	48	50 60	72 84 96 108 120

ETP	Age (months)			
	0–46	48–58	59–73	83–onwards
Treated 2-year Program				
Treated 1-year Program				
Control				
Tests	46	49 58	61 70 73	83 85 95 121 203

IHDP	Age (months)			
	0–12	12–36	37–60	61-onwards
Toddler Care+Preschool				
Toddler Care+No Preschool				
No Toddler Care+Preschool				
No Toddler Care+No Preschool				
Tests	36		60	96 216

Legend

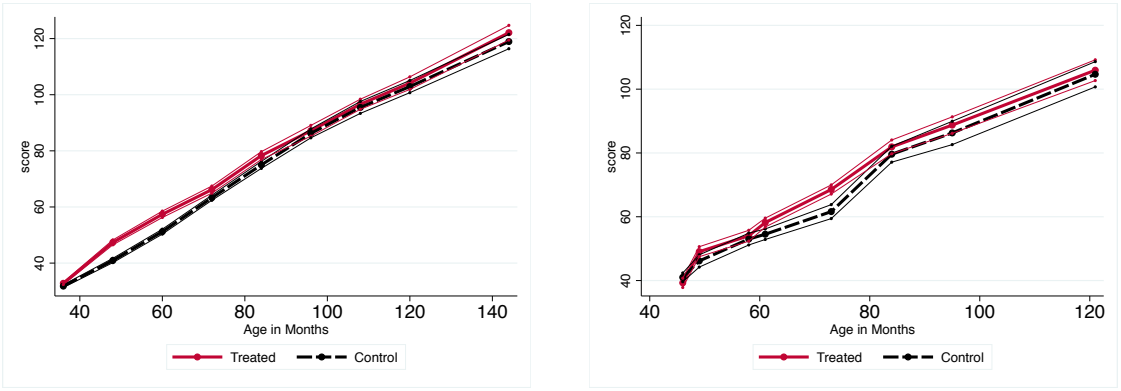
	No Education
	Early Childhood Education (ECE)
	Preschool
	School

This table presents the timing of the three programs I use in this paper. The main patterns are (i) the differences in school entry age between Perry and ETP; (ii) the difference in treatement intensity across groups of cohorts in Perry and ETP; (iii) the early age of the IHDP program; and (iv) the existence of a group that attends formal education and then stops attending in IHDP. The positioning of the tests and the length of the cells are roughly proportional to the age. Ages of program participation and test-taking in Perry and ETP were based on the school-year calendar, so they are only approximations. The ages of testing in IHDP are close to the chronological age. Educational experiences in Perry and ETP are in general deterministic functions of age and randomization status. In IHDP, educational experiences are choices made by the families, although Toddler Care is heavily influenced by the IHDP randomization, as discussed in Section 1.5.

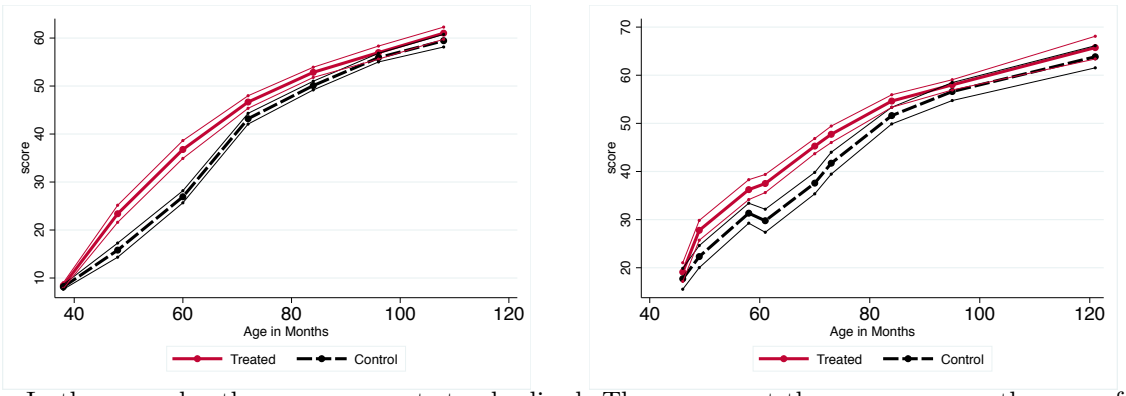
Given that $b^m > 0$, we can identify the sign of the change in skills across periods from the sign of the change in tests. Figure 1.1 presents the trends for both programs in two different tests.[21]

Figure 1.1: Dynamics of Raw IQ Scores

(a) Perry Preschool Project, Stanford-Binet (b) Early Training Project, Stanford-Binet



(c) Perry Preschool Project, PPVT (d) Early Training Project, PPVT



Note: In these graphs, the scores are not standardized. They represent the raw scores, or the sum of the number of correct questions in each year. The solid line represents the trajectory of the treatment group, and the dotted line represents the trajectory of the control group. Thin lines surrounding trajectories are asymptotic standard errors.

It is possible to see from Figure 1.1 that a gap opens when children enter the preschool programs, and it closes a few years after they have ended. More importantly, we can conclude from the figures that skills increase across all periods during childhood, with an exception for the control group of ETP around 60 months.[22] For this paper, it is important to note that

[21]The data in Figure 1.1 are original to this paper. Only age-standardized scores were originally available in the Perry data (which is usual practice in many datasets with IQ tests). Thus, in order to obtain the raw scores age-by-age it was necessary to do reverse-engineering on the scores using the original test manuals. See Appendix C for details on this process.

[22]There are only 21 individuals in that group, and the difference between measurements is only three

the treatment group does not suffer a fall in its skill level at school entry (the tests reflecting school entry in ETP are taken around 83 months). The fadeout phenomenon can, in this sense, be better described as a "catch-up" of the control group. This fact is important for the policy interpretation of the whole fadeout phenomenon. It is, however, less useful for distinguishing between the different hypotheses about the causes of fadeout. Negative trends at school entry would be evidence toward the existence of depreciation in skills (program-exit effect). However, if (i) there are positive school entry effects; or (ii) there are positive age effects, the existence of depreciation could still be compatible with positive trends in the data. Figure 1.2 addresses the first of the two problems.

1.4.4.2 Skills Increase Even for Children Transiting from Toddler Care to No Preschool

Figure 1.2 presents the evolution of raw scores for IHDP individuals that were randomized into the high-quality IHDP program, but then received no preschool education in the two years after IHDP ended (at ages four and five).[23] This is the case in which we would be most likely to see some depreciation in observational data, but the effect of age is still present.

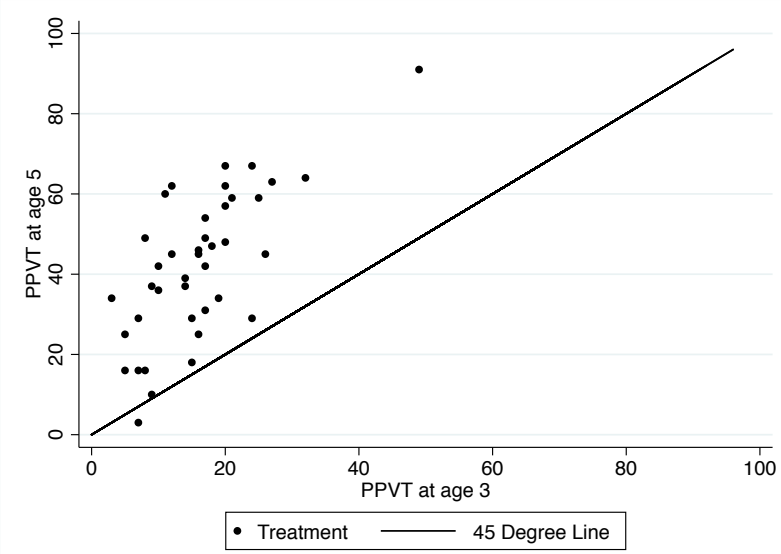
Out of 40 children randomized into treatment who received no preschool at ages four and five, only one suffered depreciation of his scores, and in that case it was a very small decrease. Although it is still possible that the fixed age effects are masking some degree of depreciation for all of the observations, I interpret this as evidence against the importance of depreciation in the data.

In Appendix F, I analyze the skill trajectories assuming that tests only have ordinal properties, using a stochastic dominance analysis and testing for equality of distributions. In Appendix G, I present trajectories using six different transformations of all the different tests available in the data. The patterns of (i) strong initial impacts; (ii) no impacts in the second year; and (iii) strong negative impacts at school age hold very generally. In Appendix

months.

[23] About two-thirds of the individuals randomized into IHDP participated on it.

Figure 1.2: Evolution of Raw IQ Scores for Children Transiting from Toddler Care to No Preschool



Note: This chart includes all children in the Infant Health and Development Program randomized into in the treatment at age 0-3, and then did not have any formal education experiences at ages four to five. Observations over the 45-degree line correspond to children who increased their scores during that period.

H, I present t-tests of difference in means for all ages for several transformations of the test scores.

1.4.4.3 Age-Standardized Scores Show Impacts of School on the Controls

As discussed in Section 1.3, under a measurement model it is possible to draw lessons about skills using test scores, if interpreted correctly. Age-standardized scores measure the skills of a child relative to the average US children of her age group.[24]

Figure 1.3 depicts the trends of age-standardized scores for the PPVT test in Perry and ETP.[25] To avoid distorting the timing of the impacts, this figure only includes cohorts that entered Perry at age 3.[26] We can see that (i) there are very strong impacts for the treated

[24]In Appendix G I present a formal definition of age-standardized scores.

[25]See Figure 48 in Appendix L for the analogous chart using the Stanford-Binet Test, which is also common for both interventions and available in the data for all years. Although that chart looks a bit more noisy, when taken together both of them share the patterns I highlight here.

[26]The graph of the Perry fadeout commonly presented in the literature mixes cohorts that entered at different times into the program. This slightly distorts the picture in a way that makes fadeout seem less pronounced than it really is.

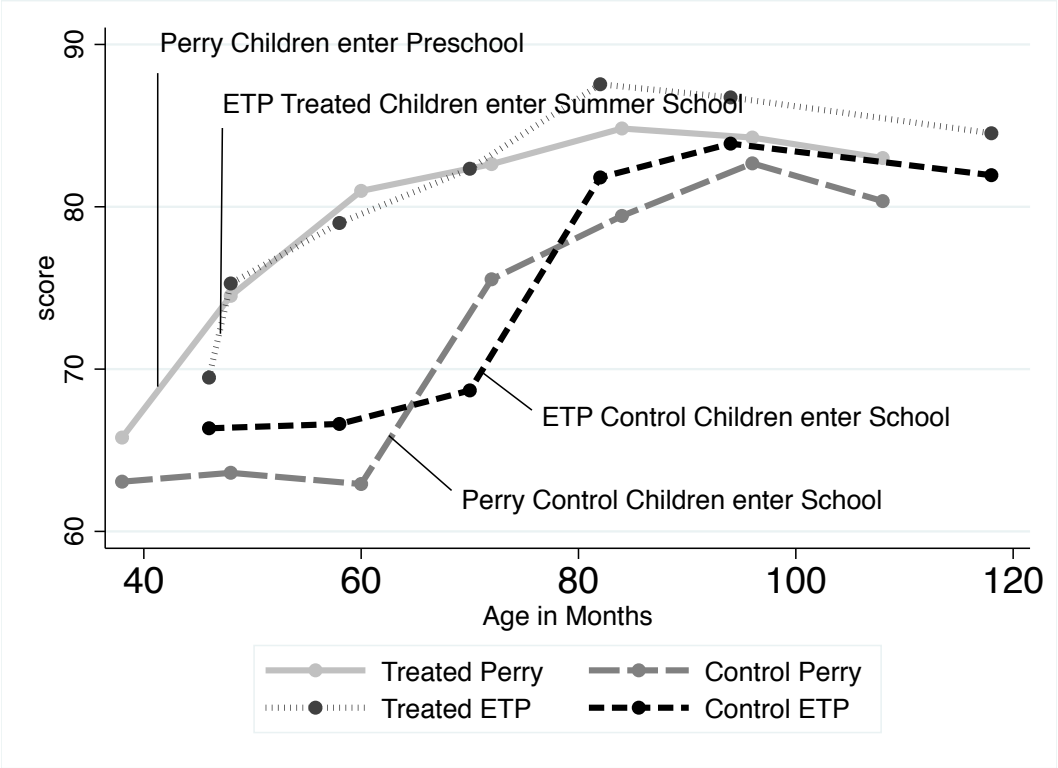
children at preschool entry; (ii) The scores of the treated children show no clear change in trend at school entry; and (iii) the fadeout seems to be strongly influenced by increases in the scores of the control group at school entry age. The surge in the scores of the control group is not an age effect: ETP children entered school at age 6, and Perry children entered at age 5. This difference creates a natural variation that allows to control for the effect of age. In Appendix L, I aggregate the data from both programs and formally test the impacts at school entry for the control group controlling for age. Across different specifications, there is a large positive impact on scores in the control group at school entry, equivalent to 3/5 of a standard deviation of that test in the population.[27]

The impacts on the control group at school entry age imply that the gap in skills between the control group (formed by very disadvantaged children) and the average US children narrowed down at that age. Appendix M formally discusses what age-standardized scores measure and how these impacts can be interpreted as evidence for the two models in Appendix B. When Perry control children entered school at age five, around 30% of US children had attended preschool education at ages 3–4. When ETP control children entered school at age six, around 70% of US children had attended Kindergarten at age 5 (Snyder et al., 2016). There might be catch-up because those previous advantages (and home environment differences) imply that control children gain more from entering schools than the average US children. On the other hand, explanations based on depreciation do not seem plausible in explaining the fadeout observed at school entry: there is no reason why the average US children at the time should have suffered a strong decrease in quality from their preschools to their schools. That implies that depreciation based on quality is not a convincing explanation. The program-exit hypotheses that explain fadeout by the supposed low quality of the subsequent schools would predict that the the scores of both groups of children should have decreased significantly compared to the average US children at school entry. This analysis suggests that an important part of the fadeout was caused by positive

[27]This regression can have a causal interpretation: entry into school is a deterministic function of age and dataset (Perry or ETP) and no treatment substitutes were available for control children in these studies.

impacts of school entry in the control group and is consistent with the evidence I present later in the paper.

Figure 1.3: Dynamics of the Standardized Scores in the PPVT Test: Perry and ETP



Note: Standardized Scores are constructed by substracting the mean for a representative sample of the age of the child to her raw score, and then dividing by the standard deviation for a representative sample for that age. The result is then multiplied by 15 and added 100, for simplicity and homogeneity across tests. Children in Perry take a baseline measurement at 36 months, are tested after the first year of preschool at 48 months, are tested after the second year of preschool at 60 months, are tested after the first year of school at 72 months and up to age 12. Treated children in ETP take a baseline measurement at 46 months, are tested after the first summer school at 49 months, receive home visits during the second year, are tested before the second summer school at 58 months, then after it at 61 months, then they receive home visits during an additional year, and they are tested before the third summer school at 70 months and after it at 73 months, when they enter school. They are tested after the first year of school at 83 and 85 months, and again at 95 and 121 months. Control children experience the same testing schedule.

1.5 Methodology

In this section, I first present the general skill formation model in which my estimations are based. Then, I discuss how to identify and estimate the period-by-period impact of the Perry and ETP programs. I use (i) the relationship between skills and test scores that was posed in Section 1.3 and (ii) the timing of events that characterize the programs, as discussed in Section 1.4. Finally, I show how it is possible to separate the program-exit and the subsequent-schooling effects in the IHDP dataset.

1.5.1 A Skill Formation Model

Given that the interventions were randomly assigned, it is possible to identify their total (accumulated) impacts on the skills in each period as the simple gap between the treated and control groups. However, I am interested in going beyond the simple treatment effects and exploiting the longitudinal structure of the data to estimate how cognitive skill gains in a given period can be decomposed from gains in previous periods. Because the programs had a single randomization, it is necessary to impose some structure to calculate this decomposition.

I start by presenting a model that nests mosts of the hypotheses in the literature. Given that this model cannot be fully identified using the available datasets, I impose some restrictions in the model to allow for identification of interpretable parameters.

Let the cognitive skill in a given period, θ_{it} be given by (i) the persistence of the cognitive skill in the previous period, θ_{it-1}; (ii) participation in formal education in period t, F_{it}; (iii) exposure to formal education in the previous period, F_{it-1}; (iv) Parental Investments in period t, I_{it}; (v) an age fixed effect, ω_t; (vi) an individual fixed effect, α_i; and (vii) a random error, η_{it}. In Appendix B, I show that the main models I consider can be reduced to these main factors. To allow for the different possibilities discussed in the hypotheses in Section 1.2, I allow for the effect of previous schooling to influence the change in skills in the current

period, both directly and interacted with current schooling:

$$\theta_{it} = \rho_t \theta_{it-1} + \tilde{\beta}_t F_{it} + \tau_t F_{it} F_{it-1} + \delta_t F_{it-1} + \tilde{\phi}_t I_{it} + \omega_t + \alpha_i + \eta_{it}. \tag{1.6}$$

The program-exit hypotheses will imply that, conditional on the level of the skills in the previous period, children that have been in formal education in the previous period will have a decrease on their skills relative to other children (so $\delta_t < 0$). An equivalent way to see this is that children in the programs already obtained positive impacts from F_{it-1}, but most of those impacts will not last.[28] I show a formal model of depreciation in Appendix B that has this implication. The subsequent-education hypotheses will imply that there is a negative effect of the interaction between previous formal education and current formal education (so $\tau_t < 0$), because children that already participated in the programs will learn less.

In practice, the main assumption I take is that the skills have complete persistence, so $\rho_t = 1$ in Equation (1.6). I assume these values instead of estimating them because it would not be possible to estimate the parameter reliably (especially while using individual fixed effects) in a dataset of this size. As discussed in Blundell and Bond (1998), and in much of the later literature on dynamic panel data estimation, autoregressive parameters that are close to unity are highly problematic to estimate. By assuming that $\rho = 1$, it is possible to estimate the rest of the parameters of the models consistently, even including fixed effects.

There are several reasons why assuming $\rho = 1$ makes sense in the case of the paper. First, Cunha and Heckman (2008a) estimate linear models of skill formation, finding that the value of the coefficient for cognitive skills ranges from 0.92 to 0.99.[29] Second, in Appendix I I present my estimations of ρ using the data from Perry, not including fixed effects, and find that for the combined cognitive measurements, and for three of the four types of tests

[28] Todd and Wolpin (2003) discuss how models in which skills depend on previous skills and contemporaneous investments implicitly assume the same decline of impacts for all inputs. By having formal education in the previous period in the model, I am allowing for the decline of the impacts for early education programs to be larger than for all other inputs.

[29] Some of the specifications in Cunha and Heckman (2008a) include fixed effects, which is an important advantage over the specifications I could estimate.

in the sample, I find estimates that cannot be statistically distinguished from 1.[30] Third, an alternative way of thinking about this assumption is that I am estimating a model in differences rather than in levels, which is a common practice in econometrics. Fourth, the main statistics I construct are period-by-period rather than asymptotic or aggregated across several periods. Thus, in practice, there is little difference between a parameter that has the value of 1 and a parameter that has a value close to 1 for my results. Fifth, as I show in Section 1.4, the evidence suggests that depreciation of the general cognitive skills is not relevant in practice. Thus, there should be no reason for the ρ parameter to be less than 1.[31]

1.5.2 Identification of Parameters in the Perry and ETP Datasets

In Perry and ETP, $F_{it-1} = F_{it-1}F_{it}$: whenever children enter formal education, they stay enrolled on it. This implies that these datasets will not allow to disentangle between some of the important hypotheses in the literature.[32] However, they will be useful on identifying the timing of fadeout.

Given that F_{it}, F_{it-1} and I_{it} are endogenous, I use the randomization to identify the parameters of Equation (1.6). I will only be able to identify certain parameters, and in certain periods of time. In my models and estimations using Perry and ETP I use differences between the treatment and control groups, Equation (1.6) becomes:

$$\Delta\theta_t - \Delta\theta_{t-1} = \tilde{\beta}_t \Delta F_{it} + (\tau_t + \delta_t)\Delta F_{it-1} + \tilde{\phi}_t \Delta I_{it}. \tag{1.7}$$

1.5.2.1 Parental Investments

Crowding out of parental investments has been discussed as a possible reason for the perceived failure of early childhood education programs (Becker, 1991). However, it has not been

[30]The estimations for Perry are implemented using ages three to ten. In the ETP data, I estimate separately because the spacing of the data is not homogeneous across measurements, and the estimates are not very stable.

[31]If ρ is greater than one, my main results, showing a strong narrowing of the gap at school entry, would be reinforced.

[32]This problem is solved using the IHDP dataset later in the paper.

show to be relevant in practice, possibly because many early childhood education programs also include parental-education components (Gelber and Isen, 2013).

Assuming that the randomization changes parental investments only through changes in participation in the programs, which is not a particularly strong assumption, it is possible to simplify Equation (1.7). In particular, I allow investments to depend on the participation in the programs,

$$I_{it} = \beta^I F_{it} + \epsilon^I_{it}. \tag{1.8}$$

But I assume that ϵ^I_{it} is not affected by the randomization ($\Delta I_{it} = \beta^I \Delta F_{it}$). To save notation, in the rest of the paper I use $\beta_t = \tilde{\beta}_t + \tilde{\phi}_t \beta^I$. Then, we obtain:

$$\Delta\theta_t - \Delta\theta_{t-1} = \beta_t \Delta F_{it} + (\tau_t + \delta_t)\Delta F_{it-1}. \tag{1.9}$$

1.5.2.2 Parameters Identified

In Perry and ETP, participation in formal education and previous exposure to it are deterministic functions of randomization and age, as there were no treatment substitutes at the time.[33] Only children randomized into the program have access to formal education before school entry, and all children have access to education after school entry: $F_{it} = 0$ before the program starts; $F_{it} = R_i$ during the program; and $F_{it} = 1$ after the program ends, and all children are enrolled in school. This implies that depending on the period, some of the parameters of the model could be directly estimated by Equation (1.9) in terms of measurements.

Identification in the model depends on transitions in F_t and F_{t-1}. There are four types of transitions in the data where the parameters of the models can be identified: (i) going from a period where no child participates in formal education to a period when treated children participate in formal education; (ii) going across two periods when only treated children par-

[33]The only exception is that in Perry some priority was given to working mothers who needed care for their children. Following Heckman et al. (2010a), I solve this problem by conditioning on the pre-treatment working status of mothers.

ticipate in formal education; (iii) going from a period when only treated children participate in formal education to a period when both groups participate; and (iv) going across two periods when both groups participate in formal education. Each of these transitions allows us to identify different parameters from the models. Table 1.2 summarizes these transitions for the case of the two-year-program cohorts in Perry and ETP. The one-year-program cases are analogous.

Table 1.2: Transitions in the ETP and Perry Data and Associated Parameters

	Formal Educ. F_t		Previous Educ. F^t		
Transition	R=1	R=0	R=1	R=0	Identified Parameters
Baseline-1st yr. Preschool	Yes	No	No	No	β_1
1st-2nd yr. Preschool	Yes	No	Yes	No	$\beta_2 + \delta_2 + \tau_2$
2nd yr. Preschool-1st yr. School	Yes	Yes	Yes	No	$\delta_3 + \tau_3$
1st-2nd yr. School	Yes	Yes	Yes	Yes	0

In Appendix B, I discuss how these parameters can be identified under a Differential Gains model and under a Depreciation model. In the Perry and ETP data, it is not possible to separate between those two models (or, more generally, between program-exit and subsequent-schooling explanations).

1.5.3 Estimation in ETP and Perry

As shown in Section 1.3, if the measurements in two periods have the same slope parameter (b^m in Equation (1.2)), a double difference in measurements across time and across treatment groups will identify the double difference in skills up to the scalar b^m. These double differences in skills will identify the main parameters of the models in Perry and ETP.

Given that, the estimation of the parameters for ETP and Perry is reduced to estimating double differences across time and across treatment groups in the measurements, controlling for the necessary covariates, and clustering at the individual level when more than one measurement is used per child. I estimate my main parameters of interest using a single regression. Let $A_t = 1$ if the observation is measured at month t after randomization and

$A_t = 0$ otherwise. Let X_i be a vector of covariates. A regression of the measurements on R_i, A_t, their interactions, and X_i will give the differences in means for each period. The difference of those estimates between each period and the previous one are my parameters of interest. The regression is clustered at the individual level to account for the fact that, in some cases, I use more than one measurement per individual per period. Then, I can obtain the simple differences for all periods using a regression for the following reduced form model:

$$M_{it}^m = \pi_0 + A_0 \cdot R_i + \sum_{t=1}^{\mathcal{T}} \{\pi_t A_t + \pi_{Rt} A_t \cdot R_i\} + X_i \phi_X + \alpha_i + \eta_{it} \tag{1.10}$$

1.5.4 Identification in IHDP

IHDP allows me to disentangle the program-exit effect from the subsequent-schooling effect because in the data $F_{i2}F_i^2 \neq F_{i1}$. However, this also brings new challenges. I use three periods in this dataset. Period 1 is ages one to three, when IHDP randomly gave access to toddler care to treated children. Period 2 is ages four to five. In this period, families decide whether to send their children to preschool or not. In Period 3, all children attend school. I use the fact that some families of children that attended IHDP decided not to send their children to preschool in Period 2 to disentangle the effects.[34] Whenever I refer to fadeout in this section, it is fadeout of the gains from being randomized to IHDP in Period 1.

For children who do not participate in preschool, there will be no subsequent-schooling effect. For that group, fadeout is a pure test of the program-exit effect. On the other hand, children who participate in preschool should suffer fadeout from both effects. In Section 1.6, I find no fadeout for the first group and complete fadeout for the second group, which I interpret as evidence of differential gains from subsequent education being the major factor in fadeout.

[34] In the second period I only consider attendance in a non-IHDP program as participating in center-based care, because IHDP ended around age 3. Thus, children could attend only for a couple of months and still report to be participating. If children attended both IHDP and a non-IHDP program, they are counted as attending.

This dataset has two characteristics that make it uniquely adequate to separate the effects: (i) IHDP had a large impact in participation in formal education in Period 1 (89% of the treated group attended a center-based program, compared to 28% for the control group); and (ii) IHDP had no significant impact in participation in formal education in Period 2 (76% of the treatment group attended a center-based program, compared to 79% for the control group). In fact, I cannot reject a t-test of identical means or a test of independence between preschool attendance and IHDP treatment status (p-value 0.29 for both). Given the relatively large size of the sample, I consider these tests as strong evidence that $F_2 \perp\!\!\!\perp R$. I now discuss how to identify the key parameters in Equation (1.6) using these characteristics of the dataset. I focus on the second period in IHDP, because it is the one that allows me to separate the different effects.

$$\theta_{i2} - \theta_{i1} = \tilde{\beta}_2 F_{i2} + \delta_2 F_{i1} + \tau_2 F_{i1} F_{i2} + \tilde{\phi}_2 I_{i2} + \alpha_i + \omega_2 + \eta_{i2} \tag{1.11}$$

Given the assumption that investments are only related to R through participation in the programs (Equation (1.8)), it is possible to decompose $E\left[I_{i2}|R=1, F_2=1\right]$ on the part influenced by the randomization, $\beta_2^I F_{i2}$, and an error that is independent from it, ϵ_{i2}^I. As before, I consider the direct impact of participation on the programs and their indirect impact through parental investments together, so $\beta_t = \tilde{\beta}_t + \tilde{\phi}_t \beta^I$. Additionally, as the error term is independent from randomization, I define $\eta_{i2}^* = \alpha_i + \omega_2 + \eta_{i2} + \epsilon_{i2}^I$.

It is not possible to simply condition on values of F_1 and F_2 to estimate the parameters, because the errors will likely be correlated with the participation decisions. In general, independent exogenous variation in both variables would be needed for the estimation of the whole model. Although I do not have exogenous variation for both F_1 and F_2, I can still estimate the key parameters by estimating the impacts of the IHDP randomization within a status of preschool attendance. This is possible thanks to $F_2 \perp\!\!\!\perp R$. I first identify the

impact of the IHDP randomization conditional on preschool attendance. I take expectations conditional on $(R = 1, F_2 = 1)$, and on $(R = 0, F_2 = 1)$:

$$\begin{aligned} E[\theta_2 - \theta_1 | R = 1, F_2 = 1] &= \beta_2 - \delta_2\beta_1 E[F_1 | R = 1, F_2 = 1] + \beta_2\tau_2 E[F_1 | R = 1, F_2 = 1] \\ &\quad + E[\eta^*_{i2} | R = 1, F_2 = 1] \\ E[\theta_2 - \theta_1 | R = 0, F_2 = 1] &= \beta_2 - \delta_2\beta_1 E[F_1 | R = 0, F_2 = 1] + \beta_2\tau_2 E[F_1 | R = 0, F_2 = 1] \\ &\quad + E[\eta^*_{i2} | R = 0, F_2 = 1] \end{aligned}$$

Given that $\eta^*_{i2} \perp\!\!\!\perp R$ by assumption and that $F_2 \perp\!\!\!\perp R$ in practice, a single-crossing model (or monotonicity assumption, as discussed in Vytlacil (2002)) is enough to remove R from the conditioning set for the errors. Let η^F be the unobservables determining distaste for preschool in a family. In principle, I allow the decision of attending preschool to be influenced by the randomization to IHDP. However, a unidirectional flow condition implies that if no impact of R on F_2 is observed in the data, then R does not change the composition of the group self-selected into preschool. Let the decision of attending preschool be given by:

$$F_{i2} = 1\left[\alpha R_i > \eta^F_i\right] \tag{1.12}$$

Let C_{η^F} be the cumulative distribution function of η^F. Given that in practice $E[F_{i2} | R = 1] = E[F_{i2} | R = 0]$, we know that $C_{\eta^F}(\alpha) = C_{\eta^F}(0)$. Thus, $\alpha = 0$ and the randomization does not affect the decision of entering preschool in any way. The conditioning sets for both groups will then be equivalent:

$$E[\eta^*_{i2} | R = 1, F_2 = 1] = E\left[\eta^*_{i2} | \eta^F_i > \alpha\right] = E\left[\eta^*_{i2} | \eta^F_i > 0\right] = E[\eta^*_{i2} | R = 0, F_2 = 1]. \tag{1.13}$$

From this result, and using both conditional expectations, I can construct:

$$\frac{E[\theta_{i2} - \theta_{i1} | R = 1, F_2 = 1] - E[\theta_{i2} - \theta_{i1} | R = 0, F_2 = 1]}{E[F_1 | R = 1, F_2 = 1] - E[F_1 | R = 0, F_2 = 1]} = -\delta_2\beta_1 + \beta_2\tau_2 \tag{1.14}$$

Doing the analogous process for $F_2 = 0$:

$$\frac{E\left[\theta_{i2} - \theta_{i1} | R = 1, F_2 = 0\right] - E\left[\theta_{i2} - \theta_{i1} | R = 0, F_2 = 0\right]}{E\left[F_1 | R = 1, F_2 = 0\right] - E\left[F_1 | R = 0, F_2 = 0\right]} = -\delta_2 \beta_1 \tag{1.15}$$

The interpretation of Equation (1.15) is clearest: it is valid to compare the means of the treated and control groups conditional on not attending preschool in Period 2, because of the independence between the conditioning variable and the randomization. Given that, the randomization provides exogenous variation in participation in formal education in Period 1. In this group, there will only be fadeout if the program-exit hypotheses are true, as in this group there is no subsequent-schooling impacts in Period 2 that could be different between groups. The interpretation for Equation (1.14) is analogous, but both effects will be present.

Given that the skills are unobserved, to estimate the expressions above in practice, I have to assume that the measurement errors in the test scores are independent not only from R, as discussed in Section 1.3, but also from F_2. This is a strong assumption, and the discussion for R on that section applies to F_2 too. I estimate the parameters for this model using two-stages least squares regressions of the difference in scores between ages five and three on the participation in formal education at ages zero to three, using the randomization as the instrument and conditioning on the participation status at ages four to five.

1.6 Results

1.6.1 Results for Perry and ETP

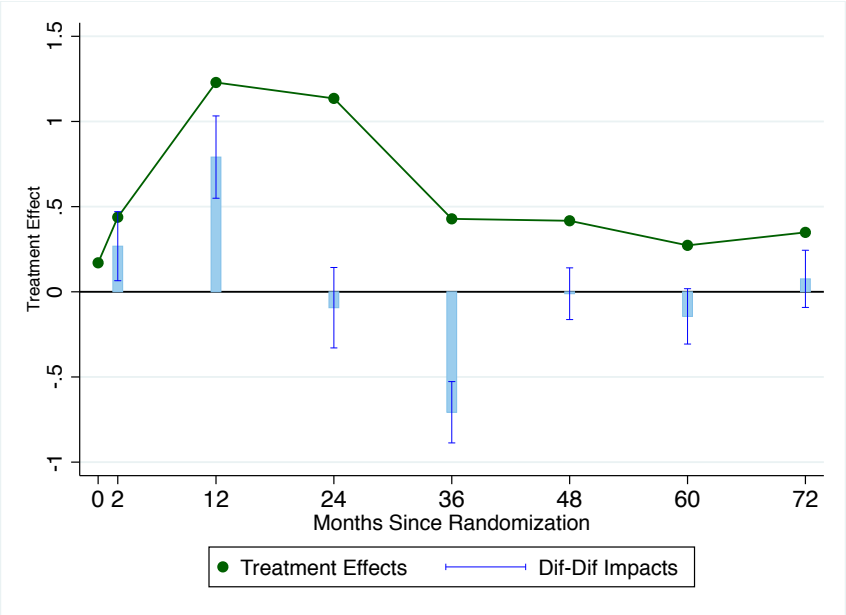
Figures 1.4 and 1.5 present the main results for Perry and ETP. The x-axes show the months after randomization for each group, which is what determines the timing of both the testing and the preschool experience in both programs. The y-axes show impacts in terms of standard deviations of the scores at each age.[35] Each chart presents two sets of results: first, the

[35] In the discussion I present the results in terms of the total fadeout, as discussed in Section 1.3

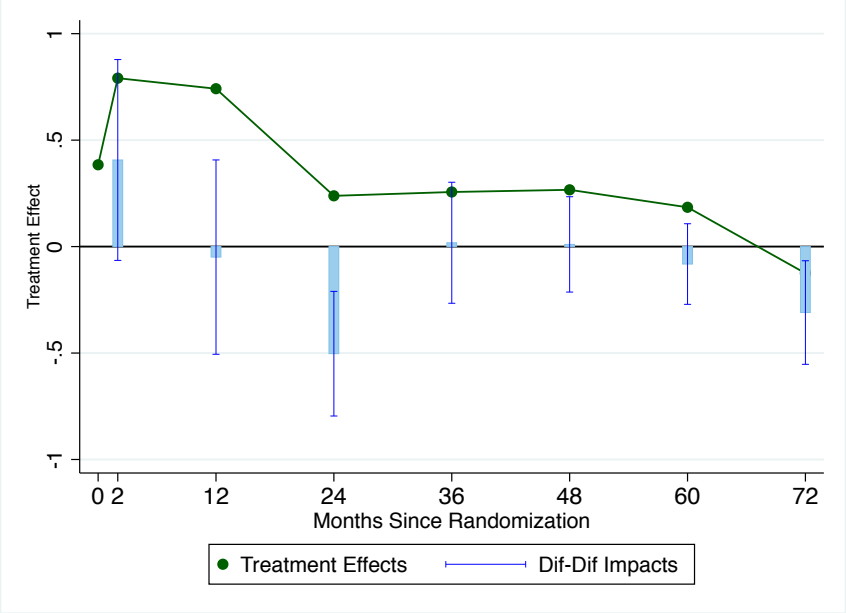
continuous line represents the evolution of the gap between the control and the treatment group. Second, the bars represent the difference between the gap in that period and the gap in the previous one. Those bars show how the events in each period affect cognitive skills, and show the main results in these figures. The thin lines around them are 90% confidence intervals. Figures 1.4a and 1.4b are based on Perry data, and Figures 1.5a and 1.5b are based on ETP data. Figures 1.4a and 1.5a are for cohorts that had two years of program, while Figures 1.4b and 1.5b are for cohorts that had just one year. Figure 1.4a is the most reliable statistically, as it is constructed from 2,574 observations for 95 individuals, while Figure 1.4b is the less reliable, as it is constructed from 660 observations for 28 individuals.[36]

[36]Figure 1.5a is constructed from 908 observations from 43 individuals and Figure 1.5b is constructed from 850 observations from 43 individuals

Figure 1.4: Differences in Differences in Perry



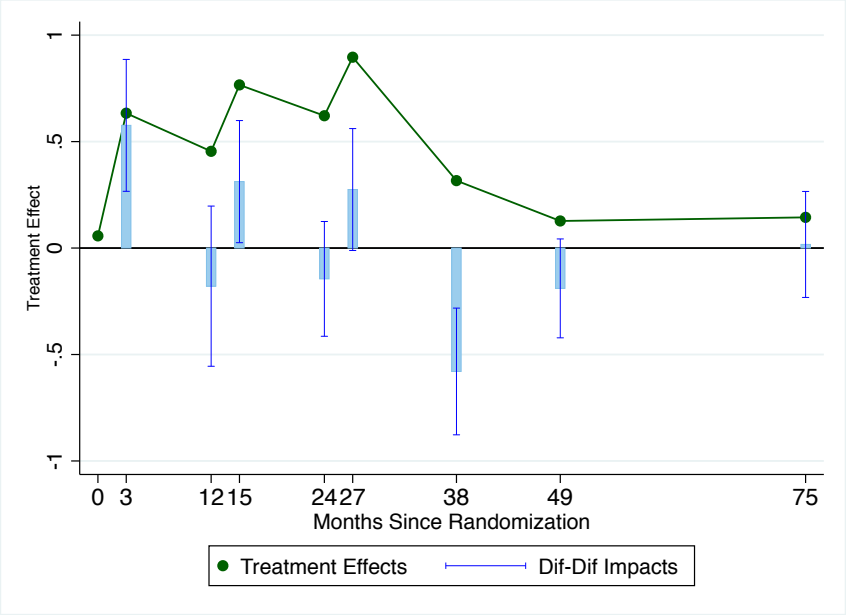
(a) Two-Years-Treatment Cohorts



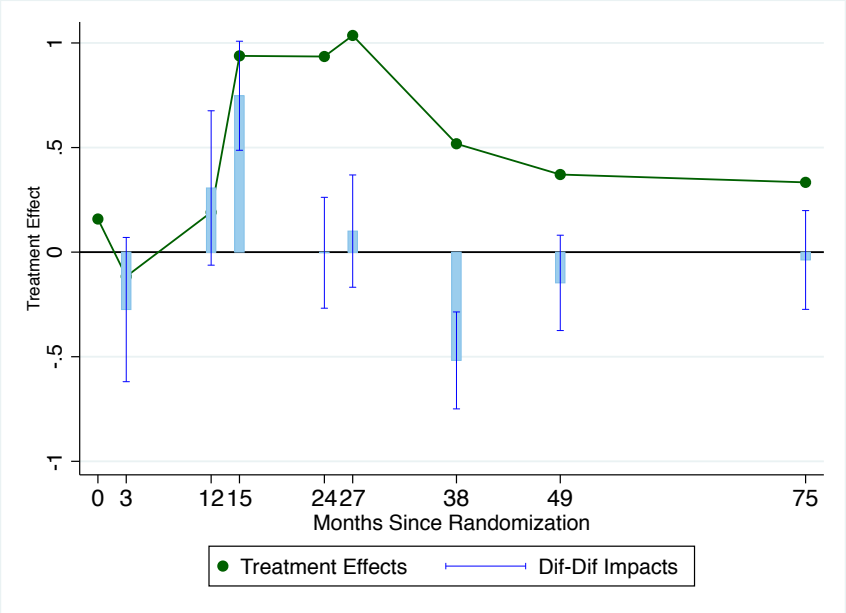
(b) One-Year-Treatment Cohort

Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. This chart uses Standardized Scores, which are constructed by substracting the mean for a representative sample of the age of the child to her raw score, and then dividing by the standard deviation for a representative sample for that age. Perry used the Third Edition of the Stanford-Binet Test. Children in Perry take a baseline measurement at 36 months, are tested after the first year of preschool at 48 months, are tested after the second year of preschool at 60 months, are tested after the first year of school at 72 months and up to age 12.

Figure 1.5: Differences in Differences in ETP



(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort

Note: The Y-axis represents impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. This chart uses Standardized Scores, which are constructed by substracting the mean for a representative sample of the age of the child to her raw score, and then dividing by the standard deviation for a representative sample for that age. ETP used the Third Edition of Stanford-Binet Test. Treated children in ETP take a baseline measurement at 46 months, are tested after the first summer school at 49 months, receive home visits during the second year, are tested before the second summer school at 58 months, then after it at 61 months, then they receive home visits during an additional year, and they are tested after the third summer school at 73 months, when they enter school. They are tested after the first year of school at 85 months, then they are tested again at 95 and 121 months. Control children experience the same testing schedule but no summer schools or home visits.

1.6.1.1 General Patterns of Impacts

1. The thin continuous lines show that the total gap between the treatment and the control group starts right after randomization, increases at the beginning of the program, is usually stable after the first year of the program, and then goes down strongly in the first year of school. After that, there might be a small difference left between the groups that might stay or disappear in the last observed periods.

2. Blue bars as early as 2–3 months into the programs show that they have significant initial impacts on scores. These impacts are remarkably strong: its magnitudes are, on average, equivalent to 63% of the total magnitude of the fadeout. These strong impacts suggest that, if schools are not completely different to preschools, it might be relatively easy for schools to make children catch up in the magnitude needed to observe the fadeout phenomenon.

3. The blue bars at 12–15 months show additional impacts of the first year of preschool. The magnitudes of the total impacts of the first year of preschool (β_1) are roughly as large as the magnitudes of the complete fadeout pheonomenon in each program.

4. The lack of significant blue bars in the second year of program shows that additional exposure to preschool has no significant impacts on scores.[37] This is a key finding because it implies that, assuming that $\beta_1 = \beta_2$, $\tau_2 + \delta_2$ is very negative, even during the programs. The interpretation of these parameters might not be obvious, as treated children are still in the program, and control children are not participating in formal education yet. In the case of the program-exit hypohteses, the only one that could explain this is that there is full depreciation of gains in each period, regardless of the quality of the educational environment.[38] In the case of the subsequent-schooling

[37] There is even a decline for 2-years group in ETP after the end of the first two Summer Schools. However, it is not significant and not replicated in the 1-year group.

[38] There is strong qualitative and quantitative evidence that the quality of the education environment in Perry was high (Kuperman, 2014; Heckman et al., 2010a).

Table 1.3: Estimates for Perry

Months Since Random	Treatment Effect 1 Cohort 1	SE	Difference in Differences Cohort 1	SE	Treatment Effect Cohort 2	SE	Difference in Differences Cohort 2	SE
0	0.17	0.10*			0.38	0.22*		
2	0.44	0.15**	0.27	0.12**	0.79	0.33**	0.41	0.29
12	1.23	0.18**	0.79	0.15**	0.74	0.24**	-0.05	0.28
24	1.14	0.17**	-0.09	0.14	0.24	0.24	-0.50	0.18**
36	0.43	0.16**	-0.71	0.11**	0.26	0.25	0.02	0.17
48	0.42	0.16**	-0.01	0.09	0.27	0.23	0.01	0.14
60	0.27	0.16*	-0.14	0.10	0.18	0.22	-0.08	0.12
72	0.35	0.17	0.08	0.10	-0.13	0.27	-0.31	0.15

Notes: this table shows estimates for experimental groups 1 (2 years of program) and 2 (1 year of program). **: significant at 5%. *:significant at 10%. Standard errors are asymptotic. All ages are in months since randomization.

Table 1.4: Estimates for ETP

Months Since Random	Treatment Effect 1 Cohort 1	SE	Difference in Differences Cohort 1	SE	Treatment Effect Cohort 2	SE	Difference in Differences Cohort 2	SE
0	0.06	0.29			0.16	0.28		
3	0.63	0.31*	0.58	0.19**	-0.12	0.34	-0.27	0.21
12	0.45	0.29	-0.18	0.23	0.19	0.30	0.31	0.23
15	0.77	0.25**	0.31	0.17*	0.94	0.26**	0.75	0.16**
24	0.62	0.29**	-0.14	0.16	0.94	0.30**	-0.00	0.16
27	0.90	0.31**	0.27	0.17	1.04	0.33**	0.10	0.16
38	0.32	0.28	-0.58	0.18**	0.52	0.29*	-0.52	0.14**
49	0.13	0.28	-0.19	0.14	0.37	0.29	-0.15	0.14
75	0.14	0.29	0.02	0.15	0.33	0.28	-0.04	0.14

Notes: this table shows estimates for experimental groups 1 (2 years of program) and 2 (1 year of program). **:significant at 5%. *:significant at 10%. Standard errors are asymptotic. All ages are in months since randomization.

hypohteses, this would imply that there are little or no skills of formal education on test scores for the treated group after the first exposure to formal education. A simple explanation for this would be the presence of large diminishing returns to education on the production of test scores. This is the most robust explanation I find in this paper, especially given the evidence that program-exit effects are irrelevant, which I present in Section 1.6.2.

5. In all four charts, there are large negative bars in the school entry year. These bars show that the relevant moment for fadeout is right at school entry. The magnitude of the negative impacts in the first year of school are comparable to the magnitudes of the total positive impacts in the first year of the programs on the treated subjects and account for 83–99% of the fadeout. As discussed before, it is not possible to separate the program-exit effects from the subsequent-education effects using these datasets.

6. Except for one case, there are no significant impacts throughout all the years after school entry.

The results are clear and consistent across the four groups. This speaks to the robustness of the results considering that, across the four graphs, there is variation in (i) the program studied; (ii) the number of years of treatment; and (iii) the ages of the children when entering school (5-6 in the case of Perry and 6-7 in the case of ETP). In Appendix J I explore the robustnes of these estimates to different types of tests and transformations of them. In most cases, the results are very similar. There are a few exceptions when using the small sample of the Perry group that only had one year of intervention (28 individuals).

1.6.2 Results for IHDP

As discussed in Section 1.5, it is possible to identify the program-exit effect from Equation (1.15) and the program-exit effect and the subsequent-schooling effects together from Equation (1.14). In this section, I show empirical estimates of those magnitudes.

I start by showing difference-in-difference estimates. First, in Figure 1.6a, I present the trend for the treatment group and the control group conditional on not having attended preschool after the end of IHDP. The change in the gap from ages 36 to 60 months (from the age IHDP ended to right before children enter schools) is free of the subsequent-schooling effect. The change in the gap for this group represents the numerator of Equation (1.15). It is the program-exit effect, scaled down because of the imperfect compliance into IHDP. Given how close this estimate is to zero, the imperfect compliance will be irrelevant in practice.[39]

Then, in Figure 1.6b, I present the trend for the treatment and the control group conditional on having attended preschool after the end of IHDP. The change in the gap from ages 36 to 60 months includes the program-exit effect and the subsequent-schooling effect.[40] The change in the gap for this group represents the numerator of Equation (1.14). It is the sum of both effects, scaled down because of the imperfect compliance into IHDP.

These statistics are easily interpretable, and give the same qualitative conclusions as the estimates corrected by imperfect compliance, which I present in Table 1.5. In that table we can see that for children that do not participate in formal education at ages four to five, having participated on formal education at ages 0 to 3 (second column) has no significant impact. The estimated impact is negative, as expected by program exit, but the magnitude is small. On the other hand, for children that participated in formal education, the impact of having participated in formal education is very strong and negative. As these estimates are in terms of standard deviations, we can see that the magnitude of the negative impact is roughly equivalent to the impacts of Perry and ETP at school entry, which further shows the robustness of these findings.

The results in IHDP allow me to separate the different models. There is no support in the data for the program-exit effect to be of a relevant magnitude in the two-years period covered by these estimates. Generalizing this result requires to assume that the fadeout before school

[39]The difference in participation between the treatment and the control group is around 70%, so the real estimate should be 1.42 times larger.

[40]In that sense, these estimates are comparable to the ones in figures 1.4 and 1.5 in Perry and ETP

Table 1.5: Estimates in IHDP

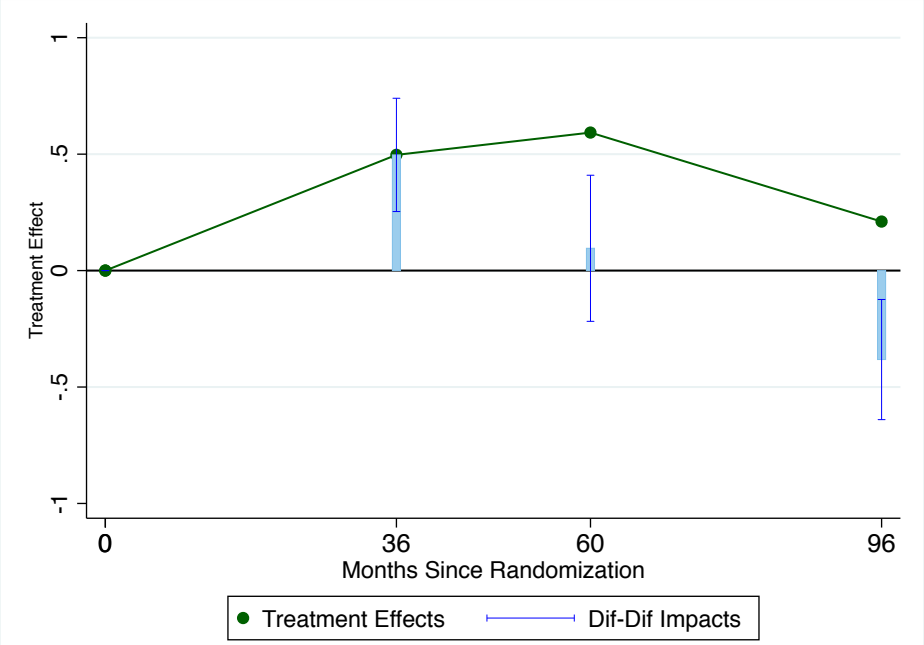
	Participation in Formal Education at Ages Four and Five	
Outcome: PPVT Std. Five - PPVT Std. Age Three	Participate	Do Not Participate
Participation in Formal Education, Ages 0 to 3	-0.76**	-0.21
	(-4.11)	(-0.94)
Constant	0.26**	-0.48**
	(2.23)	(-3.77)

Note: these estimates are obtained from two-stages least squares regessions of the growth in the standardized score of the PPVT test on participation in formal education at age zero to three, using the randomization into IHDP as the instrument. Each of the estimations conditions on a different status of participation in formal education at ages four to five, as appears in the head of the two columns. The numers in parentheses are t-statistics. ** Significance at 1% level.

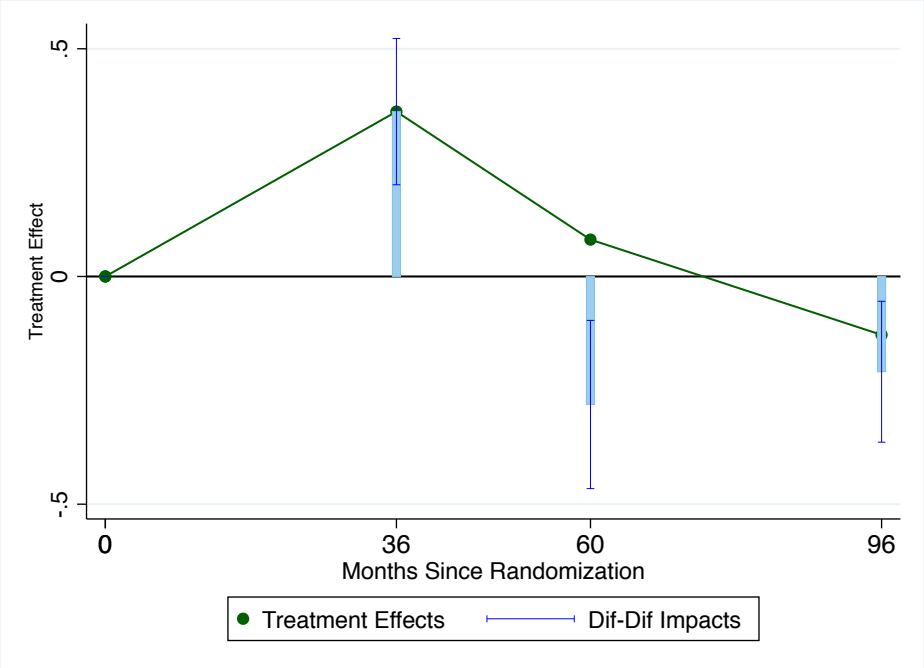
entry has similar characteristics as the fadout in schools. If that is correct, my estimates are strong evidence toward discarding program exit as being the driver of the fast convergence between the treatment and the control children in early childhood education programs. The fadeout observed in Figure 1.6b seems to be a consequence solely of differential gains from subsequent schooling. The interpretation for these results can be considered causal, given the independence between the randomization and the attendance of preschool in the second period.

Figure 1.6: Differences in Differences in IHDP

(a) Children Who Did Not Attend Preschool



(b) Children Who Attended Preschool



Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. This chart uses Standardized Scores, which are constructed by substracting the mean for a representative sample of the age of the child to her raw score, and then dividing by the standard deviation for a representative sample for that age. There is no measurement at 0 months of randomization, so the marker is not based on real data and intended to be used only as a reference.

1.7 Conclusions

In this paper I describe the fadeout of the impacts on cognitive skills of three early childhood education programs. I interpret the results by contrasting between program-exit hypotheses and subsequent-education hypotheses. Multiple pieces of evidence help to assess the empirical support of the diffferent hypotheses. First, the change in skills is almost always positive across time for all individuals. Depreciation would imply negative impacts, but we cannot discard that age effects mask those impacts. Second, the skills of control children strongly catch-up with respect to the average US children at school entry. Given that there are no clear reasons for the average US children to have negative program-exit effects, this suggests that fadeout is related to a positive impact in the control group. Third, the impacts of the programs are concentrated in the first year, and even in the first months, after preschool entry. This suggests that the large gap between the treatment and the control group is based on relatively easily trainable skills. Those skills can be gained by the control group at school entry. Fourth, there are no impacts from the second year of the programs. This helps discarding the low quality of the subsequent environments as the cause for depreciation. Fifth, 83–99% of the observed fadeout can be explained by the impacts in the year of school entry. This narrows down the possible explanations for fadeout and implies that the program-exit effect would have to be complete between consecutive years to fit the data. Finally, I use the IHDP dataset to separate the two models. I find that the role of the program-exit effect in fadeout is very small or zero. The findings in this paper are consistent: fadeout is not a consequence of the poor quality of subsequent schooling. The bulk of the fadeout is explained by school entry having a much more positive impact in the control group than in the treatment group.

There are at least two mechanisms that could explain differential gains from subsequent schooling. First, it is possible that compensating efforts from teachers help the most disadvantaged children catch-up, at the expense of more school-ready children. If this is true,

treated children would receive less investments than control children. Second, it is possible that the production function of tests scores has diminishing returns to educational investments, at least in the short run. Thus, even in absence of compensatory efforts, additional years in school will give smaller boosts to the test scores of children. The data provides one main hint about these two possibilities: The lack of additional impacts after the first year of the programs cannot be explained by the compensatory efforts explanation. However, the presence of diminishing returns is a natural explanation to this pattern that also fits well with the very large impacts of the first months of preschool.

My results are evidence for investments being dynamic substitutes at early ages. Heckman and Mosso (2014a) present a theoretical framework for dynamic complementarity and argue that the evidence is consistent with investments and endowments being direct substitutes at early ages, which implies that investments up to those ages are dynamic substitutes, as I find in this paper. Elango et al. (2015) discuss the evidence on early childhood education, finding a clear pattern of stronger impacts on most disadvantaged children.

Fadeout implies that the strong gains in cognitive skills from early childhood education programs do not give children a permanent advantage. There are at least four caveats to this statement. First, long-term exposure to stimulating environments could modify cognitive skills permanently, as research on the Flynn effect suggests (Trahan et al., 2014). These impacts are more subtle than the massive short-term impacts on test scores from the programs I study. Second, it could be possible to permanently modify the type of cognitive skills that tests measure with programs that start earlier in life (Campbell et al., 2002; Duncan and Sojourner, 2013a). Third, early childhood education programs could permanently increase other forms of cognitive skills that are not measured by the available tests (for example, the ability to come up with good answers to problems without a unique, closed solution). Fourth, there are other abilities that are improved by these programs that are valuable and mediate later gains in adult outcomes (Heckman et al., 2013a).

My findings warn against the use of short-term tests as a unique metric to evaluate

social programs. In particular, rapid increases in test scores could only be a signal of easily trainable skills that can be obtained later in life. The body of evidence in long-term impacts of early childhood education programs is still too narrow to reliably estimate a relationship between initial gains in scores and benefits in adult outcomes. For the same reason, the value of the initial increase is questionable in first place. Devoting efforts on sustaining the advantage on these skills could not be the right approach. There are advantages in other dimensions that might be more valuable than the massive, but short-lasting gains in test scores. Lerning how to measure those skills could give us a better early proxy of the real gains from early childhood education.

1.8 Acknowledgments

Thanks to Anna Ziff for excellent research assistance. Thanks to Chanwool Kim, Yu Kyung Koh, Joshua Shea, and Matthew Tauzer for their help in the construction of the data. Thanks to Dan Black, Jeanne Brooks-Gunn, Margaret Burchinal, Flavio Cunha, Steve Durlauf, Sebastián Gallegos, Jorge García, Nicolás Grau, Marianne Haramoto, Fernando Hoces, Robert LaLonde, Magne Mogstad, Derek Neal, Steven Raudenbush, Cullen Roberts, José Miguel Sánchez, Claudio Sapelli, Jeffrey Smith, Azeem Shaikh, Christopher Taber, and Petra Todd for their comments. Special thanks to Professor James Heckman for his constant support in writing this paper. The research reported in this paper was supported by a grant from Successful Pathways from School to Work, an initiative of the University of Chicago's Committee on Education. The initiative is funded by the Hymen Milgrom Supporting Organization and supported by the Division of the Social Sciences. For more information, please visit `successfulpathways.uchicago.edu` or e-mail `successfulpathways@uchicago.edu`.

Chapter 2

Analyzing the Short- and Long-term Effects of Early Childhood Education on Multiple Dimensions of Human Development

This chapter is coauthored with Jorge Luis García, James J. Heckman, Duncan Ermini Leaf, María José Prados, Joshua Shea, and Jake C. Torcasso.

2.1 Introduction

There is a growing interest in early childhood education as a means for promoting social mobility.[1] Overall state-spending on such programs increased by 12 percent in 2015. The proposed federal budget for 2017 includes a $300 million increase in spending on early childhood education.[2]

Despite the growing emphasis on early childhood education in public policy, comprehensive and methodologically rigorous evidence on its economic benefits is still scarce. Many recent studies: (i) focus on a limited set of outcomes that fail to capture a comprehensive array of program effects;[3] (ii) are based on data from follow-ups that are short-term in nature; (iii) do not correct for program attrition or for non-compliance to assigned treatment, threatening the policy-relevance of their estimates;[4] or (iv) are based on randomized controlled trials with flawed designs.[5]

Current justification for the long-term effectiveness and the efficiency of early childhood

[1] Bajaj and Labaton (2009); The White House (2014a,b).

[2] U.S. Office of Management and Budget (2015); Parker et al. (2016); Smith (2016).

[3] An extreme example is the evaluation of preschool programs using an age-eligibility cutoff. A battery of studies compare children who were just eligible and just ineligible for preschool. They therefore only assess the gains of an additional, earlier year of preschool. This does not represent a comprehensive evaluation approach; it evaluates a specific set of children for a very narrow set of tests and within a time horizon of a single year of treatment. Examples of these studies include: Gormley and Gayer (2005); Gormley et al. (2005); Weiland and Yoshikawa (2013).

[4] Consider the evaluation of Head Start through its randomized controlled trial, the Head Start Impact Study (Puma et al., 2012). Comparing subjects in the treatment and the control groups usually yields relatively low gains. This attenuation happens because a substantial proportion of subjects randomized out of the program were enrolled into preschool alternatives, some of being other Head Start centers. Thus, a raw comparison between the treatment- and the control-group subjects does not inform on either the efficiency or the effectiveness of Head Start *per se*. Studies providing a methodology to account for substitution find that Head Start has substantial effects, although they focus on a single, short-term outcome (Kline and Walters, 2015; Feller et al., 2016).

[5] An evaluation of the Tennessee Voluntary Prekindergarten is an example (Lipsey et al., 2013, 2015). The researchers designed a randomized controlled trial to evaluate the program. Unfortunately, they asked permission to assess the children after the randomization protocol. Thus, their main evaluation is based on information for children whose parents agreed for them to be evaluated *post* randomization, inducing a potential imbalance between the children randomized into and out of the program. The evaluation does not account for that. Further, results for this evaluation represent a narrow set of short-term outcomes.

education in the U.S. is largely based on evidence from the Perry Preschool Program (referred to simply as Perry). Analyses of Perry suggest that early childhood education has significant positive effects on multiple short- and long-term socio-economic outcomes, even when accounting for compromised randomization, small-sample-size inference, and multiple hypothesis testing (Heckman et al., 2010c). The analyses also show that early childhood education could have an annual internal rate of return that ranges from 7 to 10 percent.[6]

One of the criticisms of the empirical evidence favoring the economic case for early childhood education is the lack of an extensive evidence base. In response, we analyze both short- and long-term effects of early childhood education on multiple dimensions of human development using data from two randomized controlled trials, the Carolina Abecedarian Project (ABC) and the Carolina Approach to Responsive Education (CARE)—we complement this data with several non-experimental, nationally representative sources.

ABC and CARE were programs implemented in the 1970s and early 1980s. We observe short- and long-term outcomes for the subjects. The programs were separated into two phases. In the first phase, both programs randomly assigned subjects to high-quality center-based childcare from ages 0 to 5. In addition, the subjects who were assigned to center-based childcare in CARE also received home visits that aimed to foster the relationship between the subjects and their parents. Furthermore, CARE incorporated a second treatment group that received home visits without center-based childcare from ages 0 to 5. The second phase of treatment, from ages 5 to 8, consisted of home visits that aimed to continue promoting childhood development. In ABC, the second-phase treatment was randomly assigned independently of the first-phase randomization. In CARE, the second-phase was not randomized; subjects initially randomized to either of the treatment groups maintained their assignment.[7]

[6]That is, if one dollar were to be invested at age 4, and then reinvested annually and compounded over a lifetime, the return would accrue to 60 to 300 dollars by age 65. This accounts for both the program's cost and the social burden a government would cause by raising taxes to pay for it (Heckman et al., 2010d).

[7]Our main evidence is based on the first-phase component that the two programs share: high-quality

The experimental data from ABC and CARE include measures of cognitive and socio-emotional skills, educational and labor market outcomes, administrative criminal records, and a full medical examination when subjects reached their mid-30s. Data from administrative criminal records and from the full medical panel are novel to the literature evaluating early childhood education programs. The non-experimental, nationally representative data include sources to forecast life-cycle gains in public-transfer and labor income, health, and crime. Examples of these sources include: the Medical Expenditure Panel Survey (MEPS), the Medicare Current Beneficiary Survey (MCBS), and the Uniform Crime Reporting Statistics (UCRS).

Our ultimate goal is to provide a cost-benefit analysis of early childhood education programs. To construct this, we proceed in three steps. In the first step, we begin by defining the treatment-effect parameters while we estimate and state how they link to different policy questions. Our methodology accounts for different forms of attrition and non-compliance. Specifically, it considers that the parents of roughly 70% of the children randomized out of center-based childcare enrolled their children in relatively high-quality preschool alternatives. We refer to this phenomenon as control substitution.[8]

In the second and intermediate step, we provide treatment-effect estimates for a wide variety of outcomes. In doing so, a challenge arises: multiple hypothesis testing. We account for this in a standard way (Lehmann and Romano, 2005; Romano and Shaikh, 2006) while noting that it is often the case that arbitrary blocks need to be formed in order to adjust the inference using the step-down procedure. We propose and formalize an alternative: count

center-based childcare.

[8]Control substitution was not an issue in Perry. Informal conversations with Perry's staff indicate that there were no alternative preschools in the area in which subjects were treated during that time—Ypsilanti, Michigan during the 1960s. This issue is more pressing when evaluating recent programs. Examples include both ABC and Head Start—see (Puma et al., 2012) for a documentation of treatment substitution in the Head Start Impact Study.

the positive (and significant) treatment effects across the outcomes we consider. This crude summary highlights which outcome categories have the most effects, and therefore are relevant to the cost-benefit analysis, which then weighs the relative importance of each outcome.

Finally, to conduct the cost-benefit analysis, we combine the experimental and non-experimental sources of data to forecast and monetize parental income, transfer income, labor income, education, health, and crime outcomes over the life-cycle to provide estimates of the benefit-to-cost ratio and the internal rate of return of early childhood education. Because these statistics summarize the effectiveness of a program accounting for all its components in a single statistic (and a single inference test), they provide a comprehensive solution for the challenge of performing multiple hypothesis testing.

ABC's and CARE's center-based childcare from ages 0 to 5 as implemented, had substantial treatment effects on a comprehensive set of measures of human development from childhood through adulthood. For females, 78% of the outcomes we study have a *positive* average treatment effect; 31% of the outcomes we study have a *positive and significant* average treatment effect, at the 10% level. For males, the analogous figures are 78% and 29%.[9] The effects strengthen when accounting for control substitution by the families of the subjects who were randomized out of the main treatment the programs offered.

This paper extends the work of Campbell et al. (2014), who analyze the effectiveness of ABC at improving long-term health outcomes. We extend the analysis by (i) assessing multiple measures of human development; (ii) accounting for control substitution; and (iii) providing an alternative to test multiple hypotheses.[10] Furthermore, we complement the analysis by studying ABC together with CARE.

[9]These results account for program attrition.

[10]Campbell et al. (2012) also precede our work. The authors estimate treatment effects on adulthood outcomes in ABC. Unlike our approach, the authors do not assess outcomes such as health status, criminal behavior, and socio-emotional skills.

The cost-benefit analysis of ABC and CARE provide composite measures of the program's efficiency that weigh these treatment effects according to their cost to society. The pooled benefit-to-cost ratio, 4.35 (s.e. 2.57), and internal rate of return 13% (s.e. 11%), indicate that ABC and CARE are an efficient program when considering the life-cycle trajectories of the subjects.

Two previous related pieces of work provide a cost-benefit analysis of ABC (Masse and Barnett, 2002; Barnett and Masse, 2007). Their analysis is limited to outcomes up to age 21, before any of the labor income, crime, and health benefits of the program arise according to our findings. It does not provide standard errors or an analysis of the estimates' sensitivity to different modeling assumptions. Kline and Walters (2015) provide a back-of-the-envelope cost-benefit analysis of Head Start using the Head Start Impact Study. They do not analyze the life-cycle benefits and costs of early childhood education.[11]

The paper proceeds as follows. Section 2.2 provides an overview of each program. It includes a description of the eligibility criteria and the populations served, a characterization of the randomization protocol and control substitution, a comprehensive summary of the treatment, and a description of the data sources. Section 2.3 formalizes our methodology by discussing how we correct for compromised randomization and control substitution, how we test for treatment effects across multiple outcomes, and how we forecast outcomes across the life cycle. Section 2.4 presents our main results. Section 2.5 concludes. An extensive appendix presents a thorough description of the program and its costs, the data, and details on how we monetize the life-cycle outcomes. It also discusses various alternative methodologies to

[11] We present our own back-of-the-envelope cost-benefit analysis in Appendix R. It is in the same spirit to that of Kline and Walters (2015). It considers only the gains on labor income implied by the gain in kindergarten IQ proposed by Chetty et al. (2011). For simplicity, we restrict this analysis to ABC and find that the benefit-to-cost ratio is 0.47. This reinforces the idea that a comprehensive evaluation of the costs and benefits needs to consider multiple dimensions of human capital, and not only the labor income gains implied by short-term IQ.

evaluate the programs, and documents the results we present to a further extent.

2.2 Background and Data Sources[12]

2.2.1 Overview

The Carolina Abecedarian Project (ABC) and the Carolina Approach to Responsive Education (CARE) programs were designed and implemented by researchers at the Frank Porter Graham Center (FPGC) of the University of North Carolina in Chapel Hill. The programs targeted disadvantaged children from the semi-rural communities in the surrounding area.

ABC recruited four cohorts of children born between 1972 and 1977. CARE recruited two cohorts of children, one born in 1978 and one in 1979. The recruitment process for each study was identical. Potential families were referred to researchers by local social service agencies and hospitals at the beginning of the mother's last trimester of pregnancy. Eligibility was determined by a score of 11 or above on a High-risk Index (HRI).[13]

To better characterize the socio-economic status of the families participating in ABC and CARE, we construct two comparison groups using the Panel Study of Income Dynamics (PSID), a nationally representative cohort of children born in the same years as the ABC and CARE subjects (1972-1979), and a similar cohort restricted to black children. We show a comparison in Figure 2.1. Comparing the two nationally representative groups, ABC subjects were were born to younger, less educated mothers, most of whom were raising their children without the support of a father. The CARE subjects were similarly disadvantaged compared to nationally representative groups with respect to these basic household demo-

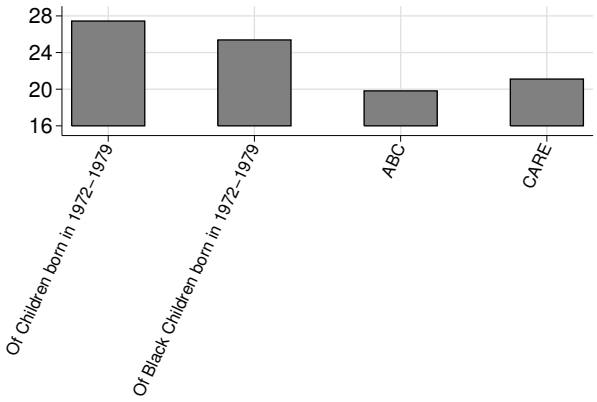
[12]This section of the paper is based on joint work with Sylvi Kuperman. We expand it in Appendix A and Appendix I.

[13]Examples of variables in the HRI are maternal education and father's stability at work. See Appendix A for details on the construction of the HRI.

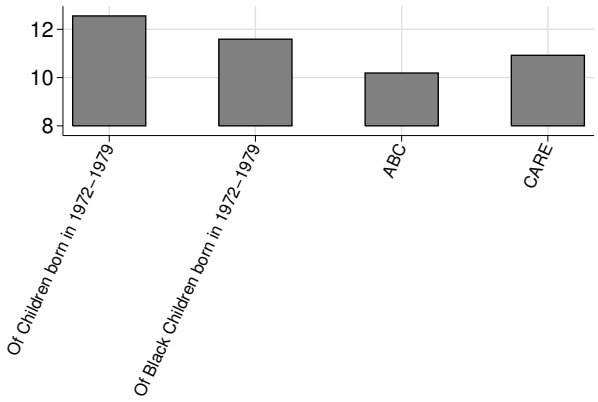
graphic characteristics.

Figure 2.1: Family Environment Baseline Characteristics, ABC and CARE

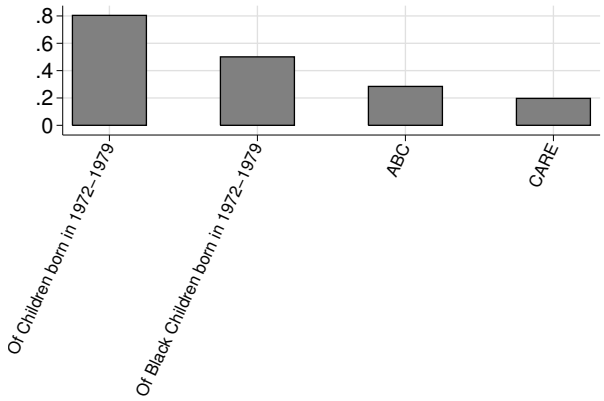
(a) Average Maternal Age



(b) Average Maternal Years of Education



(c) Proportion of Households with Father at Home



Note: These panels plot mother's age, mother's education, and an indicator of the father's presence at home. In each panel, the first bar shows the national-level for a cohort born in the same years as the ABC and CARE subjects (1972-1979), obtained from the Panel Study of Income Dynamics (PSID). The second bar uses this same information restricted to black individuals. The third and fourth bars plot the same variables for ABC and CARE, pooling the treatment and control groups.

The design and implementation of both programs were similar. Both studies had a small sample size. ABC recruited 122 subjects over four cohorts, while CARE recruited 67 subjects over two cohorts. ABC had two phases, the first of which lasted from birth until age

Table 2.1: ABC and CARE, Programs Comparison

	ABC	CARE	ABC = CARE ?
Program Overview			
Years Implemented	1972–1982	1978–1985	
First-phase Treatment	Birth to 5 years old	Birth to 5 years old	✓
Second-phase Treatment	5 to 8 years old	5 to 8 years old	✓
Recruited Sample	122	67	
# of Cohorts	4	2	
Eligibility	Socio-economic disadvantage according to a multi-factor index (see Section 2.2)	Socio-economic disadvantage according to a multi-factor index (see Section 2.2)	✓
Control			
N	54	23	
Compensation	Diapers from birth to age 3, unlimited formula from birth to 15 months	Diapers from birth to age 3, unlimited formula from birth to 15 months	✓
Control Substitution	75%	74%	

(Continue)

(Continuation)

	ABC	CARE	ABC = CARE ?
Treatment	Center-based childcare	Center-based childcare and family education	
Center-based Childcare			
N	53 (participated)	17	
Intensity	6.5–9.75 hours a day for 50 weeks per year	6.5–9.75 hours a day for 50 weeks per year	✓
Components	Instruction, medical care, nutrition, social services	Instruction, medical care, nutrition, social services	✓
Staff-to-child Ratio	1:3 during ages 0–1	1:3 during ages 0–1	✓
	1:4–5 during age 1–4	1:4–5 during age 1–4	✓
	1:5–6 during ages 4–5	1:5–6 during ages 4–5	✓
Staff Qualifications	Mixed diplomas; experienced	Mixed diplomas; experienced	✓
Family Education			
N	(not part of the program)	27	
Intensity		Home visits lasting 1 hour. 2–3 per month during ages 0–3. 1–2 per month during ages 4–5	
Curriculum		Social and mental stimulation; parent-child interaction	
Staff-to-child Ratio		1:1	
Staff Qualifications		Home visitor training	
School-age Treatment			
N	46 (participated)	39	
Intensity	Every other week	Every other week	✓
Components	Parent-teacher meetings	Parent-teacher meetings	✓
Curriculum	Reading and math	Reading and math	✓
Staff-to-child Ratio	1:1	1:1	✓
Staff Qualifications	Graduate degree and training in special education	Graduate degree and training in special education	✓

Note: This table compares the main elements of ABC and CARE, summarized in this section. A ✓ indicates that ABC and CARE had the same characteristic.

5. In this phase, children were randomly assigned to either treatment or control groups. The treatment group received: (i) center-based childcare; (ii) breakfast, lunch, an afternoon snack, iron-fortified formula for the first 15 months of life, and diapers until age 3; and (iii) medical care from licensed nurses who were supervised by a pediatrician, frequent health check-ups, and hospital referrals when serious medical treatment was needed. In contrast, the control group only received iron-fortified formula for the first 15 months and diapers until age 3. In the second phase of treatment, at the age of 5, the 95 subjects still in the study were randomly assigned again to treatment or control groups, independently of their status in the prior randomization. This second-phase treatment consisted of home visits targeting both children and parents and lasted until age 8.

CARE also had two treatment phases, though subjects were randomized only once. While the two programs had essentially identical second phases, the first phase of CARE differed from the first phase of ABC by its inclusion of a family education component. This component was designed to study the effects of improving the home environment on child development.[14] The first treatment phase of CARE lasted from birth until age 5. Children were randomly assigned to one of three experimental groups: control (23 children), family education (27 children), and both family education and center-based childcare (17 children). As in ABC, the control group received iron-fortified formula from birth to 15 months and diapers to age 3. The family education group received home visits that aimed to help parents solve common problems related to childrearing. Both treatment groups received the second phase of treatment from ages 5 to 8. The ABC and CARE programs shared many objectives and program characteristics, as summarized in Table 2.1.

The aim of this project is to evaluate early childhood education using information from both ABC and CARE. To that end, we restrict our attention to the treatment group of CARE

[14]Wasik et al. (1990).

offering the most similar programmatic content as offered to the treatment group of ABC, the center-based childcare and family education treatment group. Henceforth, we refer to this as the treatment group in CARE and do not make use of the information of the family education treatment group. For a similar reason, our main objective is to analyze the first phase of the treatment ABC and CARE offered.[15]

In both programs, from birth until the age of 8, data were collected annually on cognitive and socio-emotional skills, home environment, family structure, and family economic characteristics. After age 8, the collection of data was less frequent. Information on cognitive and socio-emotional skills, education, and family economic characteristics was collected at ages 12, 15, 21, and 30.[16] In addition, we have data that are novel to the literature evaluating early childhood education programs: long-term measures of socio-emotional skills, and administrative criminal records and a full medical panel at age 34. This allows us to study the long-term effects of the programs along multiple dimensions of human development. Table 2.2 and Table 2.3 summarize the available data. The data collection process was analogous in both programs.[17]

2.2.2 Randomization Protocol and Compromises

2.2.2.1 ABC

Both the first and second phases of randomization were conducted at the family level, so pairs of siblings and twins were jointly randomized into either treatment or control groups.[18]

[15]Separate analysis of CARE comparing the different treatment groups and comparing the family education treatment and the control groups indicate that the the family education group had very little effects across all the measures we consider. Similarly, when exploiting random assignment to second-phase treatment in ABC, we find that the second phase of treatment in ABC had little effects.

[16]At age 30, measures of cognitive skills are unavailable for both ABC and CARE.

[17]In Appendix A.3, we document the balance in observed baseline characteristics across the treatment and control groups, once we drop the individuals for whom we have crime or health information, for which there is substantial attrition. Further, the methodology we propose addresses missing information in either of these two outcome categories.

[18]Sibling pairs occurred when the two siblings were close enough in age such that both of them were eligible for the program.

Table 2.2: Data Availability (Part I)

		Early Childhood		Childhood and Adolescence		Adulthood	
Category	Sub-category	ABC Age (in months)	CARE Age (in months)	ABC Age	CARE Age	ABC Age	CARE Age
Demographics	Gender	Birth, 18, 30, 42, 54	Birth, 18, 30, 42, 54	-	-	-	-
	Race	Birth, 18, 30, 42, 54	Birth, 18, 30, 42, 54	-	-	-	-
	Birthdate	Birth, 18, 30, 42, 54	Birth, 18, 30, 42, 54	-	-	-	-
Physical Health	Growth data (e.g. height, weight)	3, 6, 9, 12, 18, 24, 36, 48, 60	Birth, 6, 12, 18, 24, 36, 48, 60	-	-	-	-
	Health issues	-	-	8, 12, 15	8, 12	21, 30	21
	Full medical sweep	-	-	-	-	34	34
Family Environment	Family Members (e.g. marital status of parents)	Birth, 6, 18, 30, 42, 54	Birth, 6, 18, 30, 42, 54, 60	6, 8, 12, 15	7, 8, 12	21, 30	30
	Family Economic Environment (e.g. parent occupations)	Birth, 18, 30, 42, 54	Birth, 18, 30, 42, 54	6,8,12,15	5,7,8,12	21	30
	Family Social Status (e.g. parents' education)	Birth, 18, 30, 42, 54	Birth, 6, 18, 30, 42, 54, 60	6, 8, 12, 15	7, 8, 12	-	-
	Physical Health of Family Members	Birth	Birth	8, 12, 15	12	-	-
	Marital Status	-	-	-	-	21, 30	21,30
	Number of Children	-	-	-	-	21, 30	30
Childcare	Childcare Experience	Birth, 18, 30, 42, 54	6, 18, 30, 42, 54	-	-	-	-
	Parental Care	6, 18, 30, 42, 54	6, 12, 18, 30, 42, 54	-	-	-	-
Cognitive Assessments	Intelligence Levels	3, 6, 9, 12, 15, 24, 15, 18, 24, 36, 48, 60	6, 12, 18, 24, 36, 48, 60	6, 7, 8, 12, 15	6, 7, 8, 12, 15	21	-
	Language Ability	36, 42, 48, 54	30, 42, 54	6, 7, 8, 12	6, 7, 8	-	-
	Motor Development	3, 6, 9, 12, 18, 24, 30, 42, 54	6, 12, 18, 24, 30, 42, 54	7	6	-	-
	Critical Thinking	30, 36, 42, 48, 54, 60, 66, 72	-	6, 7, 8	8, 12		

Table 2.3: Data Availability (Part II)

		Early Childhood		Childhood and Adolescence		Adulthood	
Category	Sub-category	ABC Age (in months)	CARE Age (in months)	ABC Age	CARE Age	ABC Age	CARE Age
Non-Cognitive Assessments	Social Skills	30, 36, 42, 48, 54, 60, 66, 72	6, 12, 18, 24	6, 8, 12, 15	8, 12	21, 30	21, 30
	Self-Control	3, 18, 30, 36, 42, 48, 54, 60, 66, 72	6, 12, 18, 24	6, 7, 8, 12, 15	12	21, 30	-
	Self-Consciousness	30, 36, 42, 48, 54, 60, 66, 72	-	8, 12, 15	8, 12		
	Work Ethic	-	-	6, 7, 8, 12, 15	6, 7, 8, 9, 12		
	Social Activities	-	-	8, 12, 15	8, 12	21, 30	21, 30
Academic Achievements	Standardized Tests	-	-	6, 7, 8, 12	6, 8, 9, 12	-	-
	Performance in School	-	-	12, 15, 17	11, 12	-	-
	Education Level	-	-	-	-	21,30	21,30
Economic Status	Living Circumstances	-	-	-	-	21, 30	21, 30
	Working Condition (e.g. job title & category)	-	-	-	-	21, 30	21, 30
	Income	-	-	-	-	21, 30	21, 30
Social Conduct	Administrative Criminal Records	-	-	-	-	mid-30s	mid-30s
	Law Breaking	-	-	15	-	21	21, 30
	Risk Taking (e.g. smoking, drinking)	-	-	-	-	21, 30	21, 30

Note: This table (Part I and II) describes the major categories of variables that were measured for ABC and CARE subjects. This is not an exhaustive list of variables, nor does it include variables from auxiliary datasets. Each age listed indicates that one or more measures of the given variable were collected from the subject at that age. Measures are collected using standardized assessments, interviews, or questionnaires.

Although we know that pairing was based on HRI, maternal IQ, maternal education, maternal age, and gender of the subject, we do not know the original pairs. The study collected an initial sample of 122 subjects. 22 subjects did not complete the first-phase of treatment as initially assigned by the randomization. We characterize each of the cases in Appendix E and document that our estimations show little sensitivity when accounting for them. We explain how we account for these cases in Section 2.3.[19]

2.2.2.2 CARE

The randomization protocol in CARE had no major compromises.[20] Of the 65 initial families, 23 were randomized to control, 25 to the family education treatment group, and 17 to the family education and center-based childcare treatment group. Two families in the family education treatment group had twins who were jointly randomized, as in ABC. There were four cases of program attrition.[21] For methodological purposes, we consider these subjects analogous to their corresponding cases in ABC. We do not present exercises to evaluate the sensitivity to non-compliance because there was none in CARE. Figure A.4 in Appendix A illustrates CARE's randomization protocol and the presence of subjects throughout the follow-ups.

2.2.3 Control Substitution

In both programs, many subjects without access to center-based childcare through random assignment attended alternative preschools. In this section, we characterize the types of

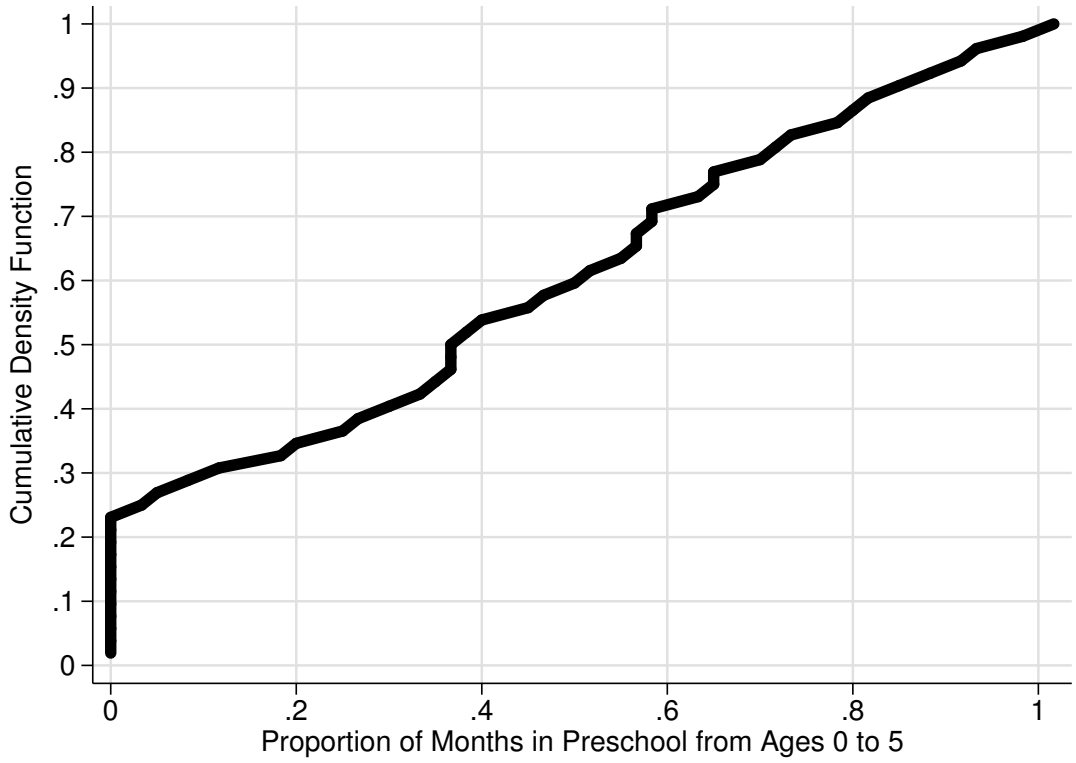
[19]In Appendix B, we compare the observed, baseline characteristics of the subjects in Table A.1 to the observed, baseline characteristics of the subjects who complied to the initial treatment assignment. We find little evidence of differences.

[20]Wasik et al. (1990); Burchinal et al. (1997).

[21]In Appendix B, we compare the observed, baseline characteristics of the subjects in Table A.2 to the observed, baseline characteristics of the subjects who complied to the initial treatment assignment. We find little evidence of differences.

care received by the treatment group. We propose a methodology to answer policy-relevant questions in Section 2.3.

Figure 2.2: Control Substitution, ABC



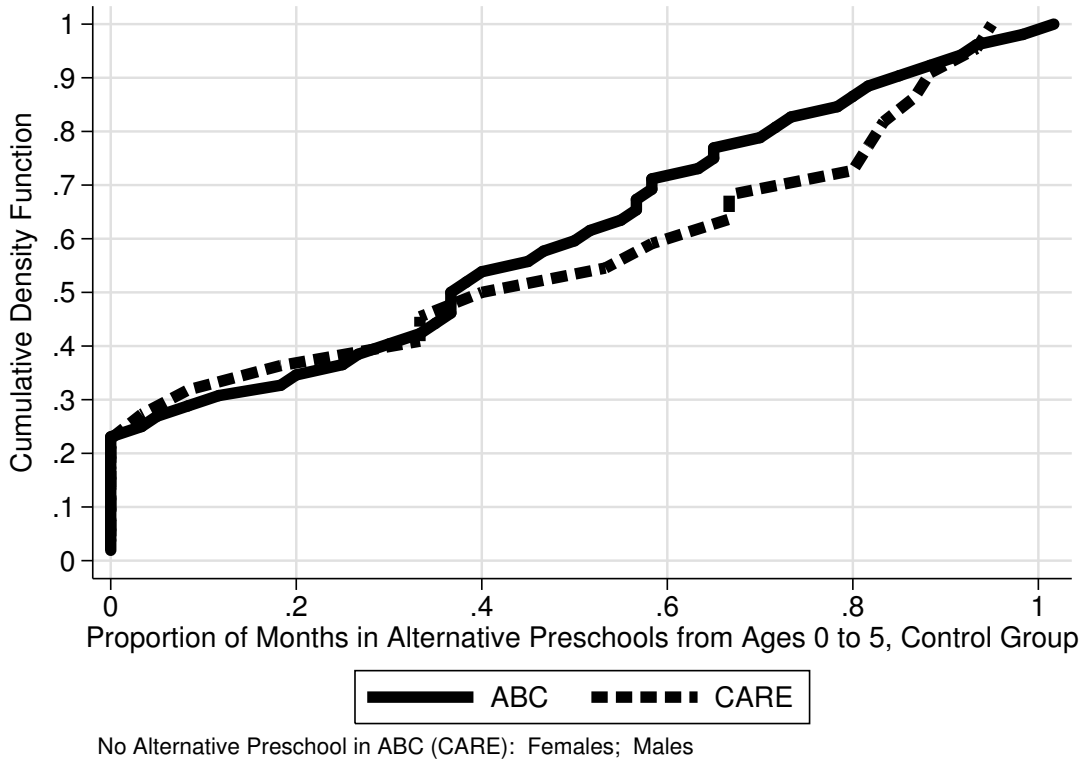
Note: This figure displays the cumulative density function of enrollment in alternative preschools of the control group in ABC.

In ABC, 75% of control-group subjects were enrolled in one of 11 local center-based childcare centers (see Figure 2.2). Most of these centers received federal subsidies and were therefore regulated by the Federal Interagency Day Care Requirements. Their staff members were required to be trained in early childhood education, and the centers were required to implement approved curricula designed to enhance cognitive, social, and linguistic competence in disadvantaged children. They had to comply to stringent staff-child ratios.[22] In CARE, 74% of the control group and 63% of the family education group were enrolled in alternative

[22]Burchinal et al. (1989).

preschools by their parents (see Figure 2.3). Parents in both of these groups had as options a similar set of local center-based childcare centers as the ABC children in the control group. We document this more thoroughly in Appendix A.

Figure 2.3: Control Substitution, CARE



Note: This figure displays the cumulative density function of enrollment in alternative preschools of the control and family education treatment groups in CARE.

2.2.4 Program Costs

The costs of the programs are a fundamental cfinput to our calculations of the benefit-to-cost ratio and the internal rate of return. We improve on previous estimates of the costs by using primary-source documentation—progress reports written by the principal investigators and related documentation we recovered in the archives of the research center where the program was implemented. We display these sources in Appendix I.

Table 2.4 breaks down the costs by different categories and items for a year of treatment. We obtain the personnel wages from the primary sources we show in Appendix I and include the type of personnel based on conversations with the programs' staff. Our sources display the actual wages without accounting for fringe benefits. We add 15% to the wages to account for fringe benefits. The costs we label as "other" account for nutrition and services that the subjects received when they were sick, diapers during the first 15 months of their lives, and transportation to the center.[23] The costs are based on sources describing ABC treatment for 52 children. We use the same costs estimates for CARE, for which information is scarcer. We have no reasons to expect sizable differences in the costs of the two programs.[24]

Table 2.4: Yearly Program Costs, ABC and CARE

Item	Yearly Cost in 2014 USD
1 Program Director	60,935
1 Social Worker	35,869
3 Lead Teachers and 2 Teachers Aides (Nursery)	204,457
4 Lead Teachers and 4 Teacher Aides (Toddlers)	305,181
2 Teaching Support Staff	53,341
1 Secretary	32,973
1 Clerk	32,537
Workers' Fringe Benefits	124,935
Other	4,891
Total	962,726
Total per Subject	18,514

Sources: See Appendix I.
Note: This table summarizes the yearly costs for ABC and CARE. They are based on primary-source documentation describing ABC. We assume that the costs for ABC and CARE were the same based on conversations with programs' staff Kuperman and Cheng (2014); Kuperman (2015).

We intend to report the cost of replicating ABC. The costs exclude research-related or

[23]The control children also received diapers during approximately 15 months, and iron-fortified formula. We assume that this generated a cost of half that amounts to half of "other" category for the first 15 months of their lives.

[24]CARE's treatment group that is relevant to our calculations, the center-based childcare and family education treatment group, received an additional service if compared to the treatment group in ABC: family education. The primary sources that we use indicate that this resulted in no additional cost. The staff implementing the center-based childcare treatment implemented the home visits without receiving an additional payment for doing so. We assume that any transportation costs to the children's home were minimal.

policy-analyses expenses. Our calculation amounts to $\$295,239$ (1979 USD). Notably, a completely independent calculation reported in Frank Porter Graham Child Development Center (1979) indicates that the yearly cost of the program was $\$275,475$ (1979 USD). Although the calculations might have been based on the same primary sources, Frank Porter Graham Child Development Center (1979) does not provide a break down of the costs.

2.2.5 Non-experimental Data Sources

Our cost-benefit analysis requires (i) *interpolation* of components we do not observe due to intermittent data collection; and (ii) *extrapolation* or forecasting of components we do not observe because the follow-ups stop when the subjects were in their mid-30s. We use multiple sources of non-experimental data representative on the national or state level to construct these interpolations and extrapolations. Table 2.5 details the components for which we do these exercises and the sources we use. Section 2.3 explains our methodology for doing so.

Table 2.5: Auxiliary Data Sources for Interpolation and Extrapolation of Life-Cycle Benefits and Costs, ABC and CARE

	Subject's Age					
Component	16–21	21–30	31–34	34–50	61–67	68–Death
Transfer Income		cNLSY	NLSY79; PSID			
Subject Income		cNLSY	NLSY79; PSID			
Health	PSID; MEPS; MCBS; HRS					
Crime	NCDPS; NJRP; NVS; UCRS					

Note: This table details the non-experimental data sources we use to interpolate and extrapolate the life-cycle benefits and costs of ABC and CARE. cNLSY: Children of the National Longitudinal Survey of the Youth 1979; NLSY79: National Longitudinal Survey of the Youth 1979; PSID: Panel Study of Income Dynamics; MEPS: Medical Expenditure Panel Survey; MCBS: Medicare Current Beneficiary Survey; HRS: Health and Retirement Study; NCDPS: North Carolina Department of Public Safety Data; NVS: National Crime Victimization Survey; NJRP: National Judicial Reporting Program; UCRS: Uniform Crime Reporting Statistics.

2.3 Methodology

2.3.1 Parameters of Interest and Policy Questions

Random assignment to treatment alone does not guarantee that conventional treatment-effect estimates commonly used in the literature are able to answer policy-relevant questions. For an estimator to be useful in policy design, it should relate to a relevant parameter by clearly stating the counterfactual scenario to which the evaluated program is being compared. We define three parameters and link them to different policy questions.

Let Ω be a set with σ-algebra $\sigma(\Omega)$ characterizing the program's subjects, with generic element $\omega \in \Omega$. Let Y denote an outcome of interest. D indicates whether or not the parents of the subject agree to participate of the program and $R|D = 1$ denotes randomization to either treatment or control status; T denotes the number of periods during the first phase of treatment—5 years.[25] We think of two counterfactuals under control status:

$Y_H^0(t,\omega)$: **Outcome under control status; subject stays at home in period** t

$Y_C^0(t,\omega)$: **Outcome under control status; subject attends preschool in period** t

We define the proportion of months in alternative preschool as

$$V(\omega) := \frac{\#\{t : Y_H^0(t,\omega) - Y_C^0(t,\omega) \leq 0\}}{T}. \tag{2.1}$$

Describing the dynamic choices underlying $V(\omega)$ is of interest but goes beyond the scope of this paper. We simplify the analysis by assuming that

[25] We define parameters that are conditional on the parents agreeing to participate of the program. That is, conditional on $D = 1$. We find little sensitivity to the few cases of non-compliance in Appendix E and adjust our estimates for the cases of attrition as we explain in Appendix G.

$$Y_H^0(t,\omega) = Y_H^0(\omega)$$
$$Y_C^0(t,\omega) = Y_C^0(\omega). \tag{2.2}$$

We write the counterfactual outcome when the child is fixed to control status as

$$Y^0(\omega) := [1 - V(\omega)]\, Y_H^0(\omega) + [V(\omega)]\, Y_C^0(\omega), \tag{2.3}$$

and make explicit its dependence on $V(\omega)$, allowing us to answer policy-relevant questions. Likewise, we write the outcome when the child is fixed to treatment status as $Y^1(\omega)$.

There are two possible approaches. One approach is to treat $V(\omega)$ as binary, where $V(\omega) = 0$ or $V(\omega) > 0$. The other approach is to allow for multiple values of V and let V to be continuous in the limit. The latter approach is ideal, because it would allow us to construct the counterfactual $Y_C^0(v,\omega)$ for $v \in [0,1]$ denoting a realization of V. This approach, however, is problematic in the context of the small number of observations in our experimental datasets. While the former approach limits the cases to either $V(\omega) = 0$ or $V(\omega) > 0$, it still allows for the definition of policy-relevant parameters. Under this approach, we can frame the parental decision in a standard Roy-type setting noting that

$$\Pr\left[Y^1(\omega) \geq \max\left(Y_H^0(\omega), Y_C^0(\omega)\right)\right] = 1, \tag{2.4}$$

where we could also frame the problem in terms of parental utility function $U(\cdot)$ over the outcome Y. We present estimates for different versions of this Roy model in Appendix D.

We focus on simpler parameters that we can directly use in the cost-benefit analysis. The estimates of these parameters and the Roy-model equivalents are qualitatively similar. The

first parameter of interest relates to the following question: what is the effect of the program as implemented? That is, what is the effect of the program without fixing the parental decision of whether or not to enroll the subject in alternative preschool? Importantly, this parameter does not speak to the effectiveness of the program by itself. Instead, it speaks to the effectiveness of the program relative to the supply of alternatives preschools that was in place when the program was implemented. The parameter is:

$$\Delta := \mathbb{E}_{\Omega}\left[Y^{1}(\omega) - \max\left(Y_{H}^{0}(\omega), Y_{C}^{0}(\omega)\right) | D = 1\right]. \tag{2.5}$$

Random assignment to either the treatment or control group allows us to identify this parameter.

It is perhaps more policy-relevant to inquire on the efficiency of a program with respect to a clearly stated counterfactual. For example, if we ask: what is the effectiveness of the program with respect to a counterfactual in which the child stays at home? A parameter associated with that question is:

$$\Delta(v = 0) := \mathbb{E}_{\Omega}\left[Y^{1}(v, \omega) - Y^{0}(v, \omega) | V = 0, D = 1\right]. \tag{2.6}$$

Random assignment to the treatment group does not identify this parameter.[26] We can approximate it with the following estimator:

$$\widehat{\Delta}(v = 0) := \widehat{\mathbf{E}}[Y | R = 1, V \in [0, \eta], D = 1] - \widehat{\mathbf{E}}[Y | R = 0, V \in [0, \eta], D = 1] \tag{2.7}$$

with $\eta \to 0$ and where $\widehat{\mathbf{E}}[\cdot]$ represents an estimate of $\mathbb{E}[\cdot]$. That is, we compare the subjects randomly assigned to treatment with the subjects randomly assigned to control in a

[26] We abuse notation to index the realization of $V(\omega)$. Differently from the definition above, the first argument in $Y^{r}(\cdot, \cdot)$ represents the proportion of time in preschool alternatives and not a time period. We do this to avoid further complicating the indices of the counterfactual outcomes.

neighborhood where subjects do not take preschool alternatives. Various matching estimators allow us to estimate how likely subjects are to take preschool alternatives, based on observed characteristics (Heckman et al., 1997, 1998). We provide different versions of these estimators below.

Similarly, we define a parameter that allows us to compare the effectiveness of the program relative to the preschool alternatives:

$$\Delta\left(v>0\right):=\mathbb{E}_{\Omega}\left[Y^{1}\left(v,\omega\right)-Y^{0}\left(v,\omega\right)|V>0,D=1\right] \tag{2.8}$$

and provide an estimate analogous to (2.7). The parameters in (2.6) and (2.8) address control substitution, in the sense that they fix the counterfactual comparison accounting for the decisions that the parents make to enroll children in alternative preschools.

2.3.2 Testing Multiple Hypotheses

We are interested in the effects that the program has on multiple dimensions of human development. We have measures of outcomes from very early in life to the mid-30s. This generates a multiple hypothesis testing problem. Two approaches are: (i) adjust the inference to account for the correlation of the outcomes using a step-down procedure (Lehmann and Romano, 2005; Romano and Shaikh, 2006); and (ii) monetize the outcomes to produce a cost-benefit analysis. We adjust the inference when estimating the parameters in Section 2.3.1 as in Lehmann and Romano (2005) and Romano and Shaikh (2006) and provide a cost-benefit analysis below. In this section, we provide an intermediate alternative that informs on the relative importance of different outcomes in the cost-benefit analysis.

Let $\mathcal{G}$ be the index set for different groups of outcomes and let $\mathcal{O}_g$ be a group of outcomes, with $g \in \mathcal{G}$. Let $F_{j,g}^{R}\left(y_{j,g}^{R}\right)$ be the marginal distribution of outcome j in group g when

randomized into treatment $R = 1$ or control $R = 0$. Assume that we want to perform inference on estimates of parameters of the type (2.5) across multiple outcomes. That is, inference on

$$\Delta_{j,g} := \mathbb{E}_{\Omega}\left[Y^1_{j,g}(\omega) - \max\left(Y^0_{j,g,H}(\omega), Y^0_{j,g,C}(\omega)\right) | D = 1\right]. \tag{2.9}$$

for the group of outcomes in $\mathcal{O}_g$. We want to test the null hypothesis

$$H_0 : F^0_{j,g} = F^1_{j,g},\ \forall\ j \in \mathcal{O}_g. \tag{2.10}$$

In practice, we test the hypothesis

$$H_0 : \Delta_{j,g} = 0,\ \forall\ j \in \mathcal{O}_g. \tag{2.11}$$

We use the following statistic to test this hypothesis:

$$T_g = \sum_{j=1}^{\#\mathcal{O}_g} \mathbf{1}\left[\widehat{\Delta}^g_j > 0\right]. \tag{2.12}$$

For inference purposes, we bootstrap this procedure and construct a null distribution. The p-value for the number of socially positive treatment effects in group g is $1 - \widehat{F}_g(T_g)$, where $\widehat{F}_g$ is the empirical bootstrap distribution of group g.[27]

A particular case is to count the positive treatment effects in the outcomes across all the groups indexed in the set $\mathcal{G}$. This allows us to avoid (i) arbitrarily picking outcomes that have statistically significant effects—"cherry picking"; or (ii) arbitrarily blocking sets of outcomes to correct the p-values when accounting for multiple hypothesis testing.

[27]For the case where we count the number of positive and significant outcomes, we use a "double bootstrap" to produce an inference on the count. We resample B_0 times to obtain the p-value for testing the hypothesis of interest for each individual outcome. This allows us to compute the number of positive and significant treatment effects, for example. We repeat this procedure B_1 times to obtain a distribution for this count. Thus, the double bootstrap consists of $B_0 \times B_1$ data resamplings.

We provide inference on this count and on a count of treatment effects that are both positive and significant for which the inference is analogous. We also provide counts for the parameters that account for control substitution.

2.3.3 Forecasting and Monetizing Life-cycle Costs and Benefits

In this section, we explain our strategy to interpolate and extrapolate the life-cycle costs and benefits of labor income, crime, and health. The methodology for doing this exercise for parental and public-transfer income is analogous to that of labor income so we suppress it for brevity. More methodological and practical details are in Appendix H, in which we also explain a solution for cases of attrition when producing interpolations and extrapolations. Based on our forecasts, we estimate the parameters in Section 2.3.1 to perform the cost-benefit analysis of the program with and without accounting for control substitution.

2.3.3.1 Labor Income

We observe labor income at ages 21 and 30. To construct a life-cycle profile, we interpolate between ages 21 and 30 and extrapolate from ages 31 to 67, in which we assume that the subjects retire. For simplicity, we suppress D and drop individual and time subscripts. Recall that R indicates whether the subject was randomized to the treatment group ($R = 1$) or to the control group ($R = 0$), conditional on having agreed to participate in the program ($D = 1$). Y is the outcome for which we want to produce a forecast—interpolation or extrapolation. In this case, the outcome is labor income. X is a vector of observed characteristics, possibly affected by the treatment—e.g. lagged values of Y; W is a vector of baseline characteristics—e.g. race and gender; S indicates whether we observe Y in the experimental sample ($S = 1$) or an auxiliary, non-experimental data source ($S = 0$).

Our objective is to recover a forecast for Y of the type

$$\widehat{Y} = \widehat{\phi}\left(R, X, W, S = 1\right) + \widehat{\varepsilon}, \tag{2.13}$$

where $\phi\left(R, X, W, S\right) := \mathbb{E}\left[Y | R = r, X = x, W = w, S = s\right]$ and $\widehat{\varepsilon}$ is a forecasting error. That is, we assume that the outcome of interest is an additively separable function of the known objects R, X, W, S and an unobserved component ε:

Assumption 2.3.1 *(Additive Separability of the Outcome)*

$$Y = \phi\left(R, X, W, S\right) + \varepsilon. \tag{2.14}$$

Identifying $\phi\left(R, X, W, S\right)$ requires three assumptions. First, the forecast is based on observed characteristics, X. Thus, we require the auxiliary datasets to share the support on observed characteristics with the experimental dataset:

Assumption 2.3.2 *(Common Support Between the Experimental and Auxiliary Datasets)*

$$\sup\left(X | S = 1\right) \subseteq \sup\left(X | S = 0\right). \tag{2.15}$$

Second, we assume that we are able to summarize the impacts that the treatment has on the outcomes with observed characteristics, given that we are not able to observe R in the auxiliary dataset. Similarly, we need to be able to summarize the difference between the individuals in the experimental datasets and those in the auxiliary datasets based on observed characteristics. This is the third assumption. The second and third assumptions are related, as they establish the requirements for being able to "link" the individuals in the auxiliary and experimental datasets when producing the forecasts. Formally, let * denote variables we do not observe. In the auxiliary dataset we have: $(S = 0, Y, X, W, R^*)$. In the experimental dataset we have: $(S = 1, Y^*, X, W, R)$. The second and third assumptions are:

Assumption 2.3.3 *(Conditional Independence and Sufficiency of S, X, W to Describe Treat-*

ment)

$$\mathbb{E}\left[Y|R=r,X=x,W=w,S=s\right]=\mathbb{E}\left[Y|X=x^r,W=w,S=s\right] \tag{2.16}$$

where x^r is a draw from the distribution of $X|R=r$.

Assumption 2.3.4 *(Conditional Independence and Sufficiency of X, W, R to Describe Presence in a Dataset)*

$$\mathbb{E}\left[Y^*|R=r,X=x,W=w,S=1\right]=\mathbb{E}\left[Y|R^*=r,X=x,W=w,S=0\right]. \tag{2.17}$$

These three assumptions imply that

$$\phi\left(R,X,W,S=1\right)=\mathbb{E}\left[Y|X=x^r,W=w,S=0\right] \tag{2.18}$$

where $\mathbb{E}\left[Y|X=x^r,W=w,S=0\right]$ is a moment in the auxiliary dataset. The estimation of $\mathbb{E}\left[Y|X=x^r,W=w,S=0\right]$ produces a residual of the form $\widehat{\varepsilon}:=Y-\widehat{Y}$ for each individual. The forecast for each individual outcome consists of $\widehat{\phi}\left(\cdot\right)$ and a draw from the empirical distribution of $\widehat{\varepsilon}$, which we call forecast error. We account for it when interpolating and extrapolating the crime and health outcomes in addition to income.

2.3.3.2 Crime

In this section, we explain how we quantify the benefits of the program from reductions in the subject's criminal activity. Two previous studies consider the impacts of ABC on crime: Clarke and Campbell (1998) use administrative crime records up to age 21, and find no significant differences between the treatment and the control groups. Barnett and Masse (2007) mention crime in their cost-benefit analysis, but they cite the previous study to claim that there are no savings coming from a reduction in crime. We consider richer data than the previous studies, which allows us to consider crime with a comprehensive life-cycle perspective: we gather various data sources, including administrative data on individual criminal

records up to age 34, and project crimes until age 50 using prediction models based on local microdata.

We consider the following types of crime: arson, assault, burglary, fraud, larceny, miscellaneous (which includes traffic and non-violent drug crimes), murder, vehicle theft, rape, robbery, and vandalism. We use data from: (i) administrative youth arrests datasets, gathered for the age-21 follow-up; (ii) administrative adult arrests datasets, gathered around age 34; (iii) administrative sentences datasets, gathered around age 34; and (iv) self-reported adult crimes datasets, gathered in the age-21 and age-30 subject interviews. Because none of these data sources capture all criminal activity, it is necessary to combine them to more completely approximate the crimes the subjects committed. These datasets are discussed more extensively in Appendix K. The data are comprehensive and cover the full potential criminal career of subjects up to age 34, including details on the types of crimes and their timing, as well as projected and effective sentences.

We use several auxiliary datasets to construct national arrests-to-sentences and victims-to-arrests ratios: (i) the National Crime Victimization Survey (NCVS) to estimate the number of victims of crime; (ii) the National Judicial Reporting Program (NJRP) to estimate the number of sentences; and (iii) the Uniform Crime Reporting Statistics (UCRS) to estimate the number of arrests. Finally, we use microdata from the North Carolina Department of Public Safety (NCDPS) to estimate a prediction model for future crimes. This dataset contains information since 1972 on every individual who was convicted of a crime and entered the state prison system.

We follow four steps to estimate the costs of crime. We summarize the steps here and present a broader discussion in Appendix K.

1. *Count arrests and sentences.* We start by counting the total number of sentences for each individual and type of crime (robbery, larceny, etc.) up to age 34. Then, we match the data on adult arrests, juvenile arrests, and self-reported crimes, to construct the total number of arrests for each individual and type of crime up to that age.[28] About 10% of the ABC and CARE samples have missing arrest data. For these cases, we impute the number of arrests by multiplying the number of sentences for each type of crime by the national arrests-to-sentences ratio for the respective crime.

2. *Construct predictions.* Based on the sentences observed before age 34, we predict the sentences that the ABC and CARE subjects will have after that age. The NCDPS data provide lifetime sentences of individuals in North Carolina, the same state in which the program was implemented. In that dataset, we estimate linear prediction models for each type of crime in which sentences after age 34 are the outcomes, and sentences up to age 34 are the regressors. Then, we apply these models to the ABC and CARE data. The outcome for each crime type is the number of future sentences for each subject, up to age 50. We assume that individuals with no criminal records before age 34 commit no crimes after age 34. We then add these estimates to the original number of sentences, getting an estimate of the lifetime sentences. To the best of our knowledge, no prior study on the benefits and costs of an educational program has used microdata to estimate a predictive model for future crimes. The predictions are an important component of total crime, as adding them increases the total count of crimes by 30%–50%. The prediction models we estimate and the results in terms of additional crimes are presented in Appendix K.

3. *Estimate number of victims of crimes.* We observe crimes that resulted in consequences in the judicial system (i.e. crimes that resulted in arrests, sentences, or both). However,

[28] In practice, we count all offenses (an arrest might include multiple offenses). This gives the correct number of victims for our estimations. The youth data have coarser categories than the rest of the data, so we assume that all property crimes were larcenies and that all violent crimes are assaults. In our sample, assault is the most common type of violent crime, and larceny/theft is the most common property crime.

it is possible that for any subject for whom we observe to have committed a crime, he committed more crimes that we do not observe. Victimization inflation (VI) is a method to capture benefits of crime reduction for crimes without consequences in the judicial system that are unobserved in the ABC and CARE data. Previous papers using this method include Belfield et al. (2006) and Heckman et al. (2010d). We start by constructing a VI ratio, which is the national ratio of victims-to-arrests for each type of crime.[29] Then, we estimate the number of victims for each type of crime committed by ABC and CARE subjects as their total arrests multiplied by the VI ratio. Additionally, we can calculate an analogous estimate of the number of crime victims using sentences, based on the VI ratio and the national arrests-to-sentences ratio. Both estimates are very similar, as shown in Appendix K. To improve precision, the estimates in the rest of our paper are based on the average of the two.

4. *Find total costs of crimes.* We use the estimates of the cost of crimes for victims from McCollister et al. (2010) to impute the total victimization costs (see Appendix K for details on the costs we use). For crimes having arrests, sentences, or both, we consider judicial system costs as well, such as police costs.[30] Finally, we construct the total costs of incarceration for each subject using the total prison time and the cost of a day in prison.

2.3.3.3 Health

We use an alternative methodology for health-related outcomes. There are three main reasons for this: (i) health outcomes such as diabetes or heart disease are absorbing states; (ii) health outcomes are highly interdependent within and across time; and (iii) there is no evident time period available to finish accounting for benefits and costs. For example, for

[29]We assume that each crime with victims is counted separately in the national reports on arrests, even for arrests that might have been motivated by more than one crime.

[30]To be able to assign costs to each type of crime, we assume that the cost of the justice system depends on the number of offenses of each type, rather than on the number of arrests. While this could very slightly overestimate justice system costs, the costs only represent about 5% of the total crime costs.

income we extrapolate up to the retirement age of 67. However, for health, we need to predict an age of death for each individual. Thus, using the notation so far, it is not sufficient to condition on W, Z, X to recover a credible estimate of the treatment effect. Instead, we use an adaptation of the Future America Model (FAM) that projects health outcomes from the subjects' early- to mid-30s up to their projected death (Goldman et al., 2015).[31]

We provide a brief description of the model in this subsection. Appendix N provides a thorough discussion. The methodology has six steps: (i) estimate the age-by-age health state transition probabilities using the Panel Study of Income Dynamics (PSID); (ii) match these transition probabilities to the ABC and CARE individuals based on observed characteristics; (iii) estimate quality-adjusted life year (QALY) models using the Medical Expenditure Panel Survey (MEPS) and the PSID; (iv) estimate medical cost models using the MEPS and the Medicare Current Beneficiary Survey (MCBS), allowing estimates to differ by health state and observed characteristics; and (v) predict the medical expenditure and QALYs that correspond to the simulated individual health trajectories.[32].

Our microsimulation model starts the health predictions at age 30, with the information on observed characteristics available at this age. We restrict it to the individuals for whom we have information from the age-34 health follow-up. This allows us to account for components that are crucial for predicting health outcomes, such as the body mass index (BMI). The models predict the probability of being in any of the states in the horizontal axis of Table 2.6 at age $a+1$ based on the state at age a, which is described by the vertical axis of the table. The crosses indicate if being in a health a state at age a is relevant for the estimation of the probability of being in a health state at age $a+1$.[33] Absorbing states are an exception.

[31]The simulation starts at the age in which we observe the subject's age-34 health follow-up. On average this happened at age 34 for both males and females, but there is variation ranging from age 30 to age 37.

[32]As an intermediate step between (i) and (ii), we impute some of the variables used to initialize the FAM models (see Appendix N

[33]In practice, the predictions are based on two-year lags, due to data limitations in the auxiliary sources we use to simulate the FAM. For example, if the individual is 30 (31) years old in the age-30 interview, we

For example, heart disease at age a does not enter in the estimation of transitions for heart disease at age $a+1$ because it is an absorbing state: once a person has heart disease, she carries it through the rest of her life. The same is true for chronic or permanent conditions such as hypertension, having a stroke, etc.

At each age, once we obtain the transition probability for each health outcome, we draw a Monte-Carlo simulations for each subject. Thus, each simulation depends on each individual's health history and on their particular characteristics. For every simulated trajectory of health outcomes, we predict the lifetime medical expenditure using the models estimated from the MEPS and the MCBS. We then obtain an estimate of the expected lifetime medical expenditure by taking the mean of each individual's simulated lifetime medical expenditure. The same procedure is applied to QALYs.

A quality-adjusted life year (QALY) reweighs a year of life according to its quality given the burden of disease. A QALY of 1 denotes a year of life in the absence of disease (perfect health). The value of QALY for an individual in a given year is smaller than 1 when there is positive burden of disease, as worse health conditions imply lower QALYs.[34] We compute a QALY model based on the EQ-5D instrument, a widely-used Health-related Quality-of-life (HRQoL) measure, available in MEPS. We then estimate this model from the PSID. Appendix N provides more details on this estimation strategy.

simulate the trajectory of her health status at ages 30 (31), 32 (33), 34 (35), and so on until her projected dead.

[34]When an individual dies, her QALY equals zero. It is worth noting that there are extreme combinations of disease and disability that may generate negative QALYs, although this case is unusual.

Table 2.6: Health State Transitions, Age a as Predictor of Age $a+1$

Age a	Age $a+1$													
	Heart Disease	Hyper-tension	Stroke	Lung Disease	Diabetes	Cancer	Disability	Mortality	Smoking	Obesity	Health Insurance	DI Claim	SS Claim	SSI Claim
Heart Disease			×				×	×	×		×	×	×	×
Hypertension	×		×				×	×			×	×	×	×
Stroke							×	×			×	×	×	×
Lung Disease							×	×	×		×	×	×	×
Diabetes	×	×	×				×	×	×		×	×	×	×
Cancer			×				×	×			×	×	×	×
Disability							×	×			×	×	×	×
Smoking	×	×	×	×	×	×	×	×	×		×	×	×	×
BMI	×	×	×	×	×	×	×		×	×				
Physical Activ.	×	×	×	×	×	×	×		×					
Binge Drinking								×	×					
DI Claim											×	×	×	×
SS Claim											×		×	×
SSI Claim														×

Note: This table illustrates how health outcomes at age a predict health outcomes at age $a+1$. The crosses indicate if we use the age a outcome to predict the age $a+1$ outcome. DI Claim: disability insurance claim; SS Claim: social security claim; DB Claim: disability benefits claim; SSI Claim: supplemental security income claim. The age a states that do not predict themselves at age $a+1$ are absorbing states by construction.

We estimate three models of medical spending: (i) Medicare spending (annual medical spending paid by parts A, B, and D of Medicare); (ii) out-of-pocket spending (medical spending paid directly by the individual); and (ii) all public spending other than Medicare. Each medical spending model is estimated through pooled weighted least squares regressions that include a persons demographics, economic status, current health, risk factors, and functional status as explanatory variables. The MCBS-based medical spending models also include lagged health because of the length of time for which MCBS subjects are observed.

Medical Expenditure before Age 30

We combine the MEPS and retrospective information in the ABC and CARE interviews at ages 21 and 30 related to hospitalizations at ages 12 and 15 and births at age 15. In addition to this retrospective information, we use use individual and family demographics to predict medical expenditure models for each age, as summarized in Table 2.7.

Table 2.7: Health Expenditure Models by Age Group, before Age 30

Explanatory variable	Age Group			
	8-11	12-14	15-20	21-30
Race/ethnicity	✓	✓	✓	✓
Education	×	×	×	✓
Asthma Diagnoses	✓	✓	✓	✓
Hospital stays	if $\geq$ 1 week	any stay	any stay	×
Births	×	×	✓	✓
Mother present	×	✓	✓	×
Father present	✓	✓	×	×
Number of siblings	✓	✓	×	×
Foodstamps	✓	✓	✓	✓
Living arrangements	×	×	✓	✓
Working, if working age	×	×	✓	✓

Note: This table summarizes the explanatory variables included in the models we use to predict medical expenditure for each age group. Possible living arrangements are: living with parents, away at college, married, or other.

The first level of each model predicts the likelihood that the subject incurred any medical

expenditure in the period. The second level predicts the medical expenditure for those with positive expenditures.

2.4 Results

2.4.1 Treatment Effects

We consider 95 measures of human development across the life cycle and count the measures for which the program had a "socially positive" effect, without accounting for control substitution.[35] We do this for both ABC and CARE by focusing on the first phase of treatment to compare subjects who received center-based childcare to control-group subjects—noting that assignment was random.[36] Figure 2.4 displays the results from this exercise: ABC and CARE positively impact a large percentage of the outcomes we consider.[37]

We can further decompose the counts in Figure 2.4 into arbitrary categories. To economize space, we present this exercise pooling ABC and CARE. That is, we decompose the effects described in the last two bars of Figure 2.4. Figure 2.5 and Figure 2.6 present this exercise. This helps us better understand the type of outcomes the programs affected. The results indicate that a large and precise fraction of effects are positive for outcomes spanning the life cycle, from parental income to crime and including a wide variety of health categories.

Next we present an overview of outcome-specific results. Appendix C displays an extensive summary of the estimates for the 95 outcomes we consider. Note that: (i) we arbitrarily

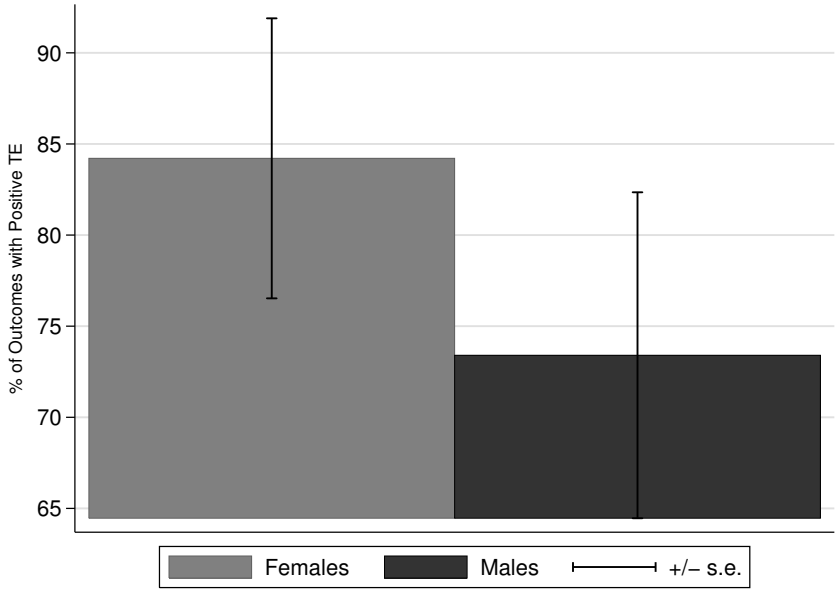
[35]These outcomes directly relate to the categories we monetize in the cost-benefit analysis. We analyze a more thorough list of outcomes in Appendix O. The results weaken, but not to a great extent. This is a consequence of adding outcomes for which it is not clear that treatment should have a positive treatment effect.

[36]In ABC, this implies comparing the subjects who were randomly assigned to the treatment group to the subjects who were randomly assigned to the control group, in the first phase. In CARE, this implies comparing the subjects who were randomly assigned to receive center-based childcare and family education to the subjects who were randomly assigned to the control group, in the first phase as well.

[37]The calculation of the standard errors follows from the bootstrap procedure we discuss in Section 2.3.

pick the outcomes we discuss next because we consider them of economic interest; and (ii) Appendix C displays results accounting for multiple hypothesis testing as in Lehmann and Romano (2005) and Romano and Shaikh (2006). We do not lose significance in the majority of outcomes.

Figure 2.4: Positively Impacted Outcomes, ABC and CARE

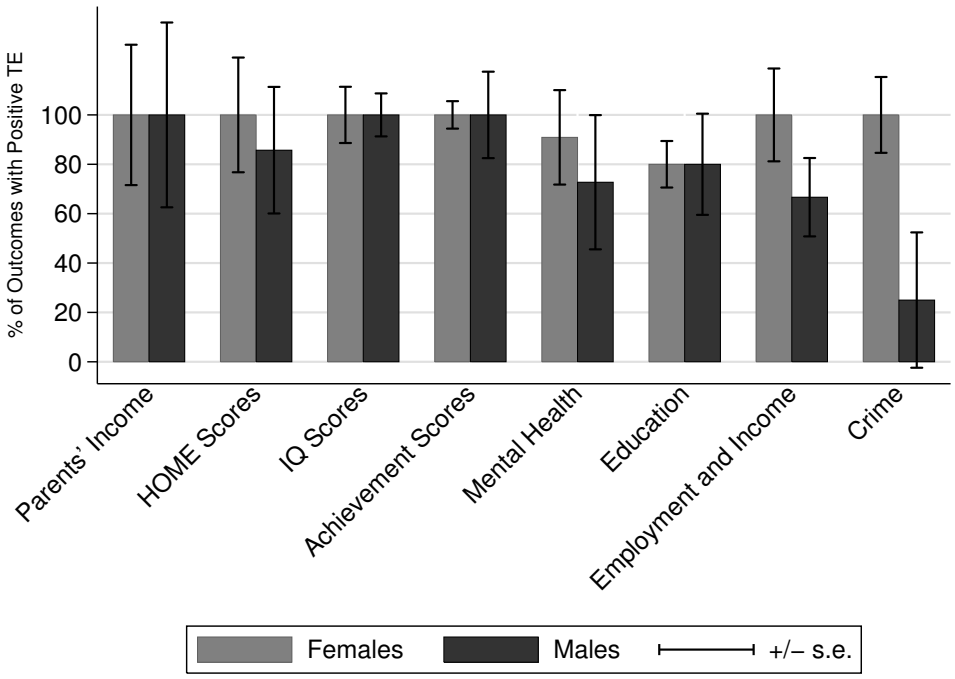


Note: The bars compare the mean of positive impacted outcomes between subjects in ABC and CARE who received center-based childcare and family education and subjects who receive either family education or no treatment at all.

Table 2.8 presents the results for females. We focus on three columns. Columns (2) displays the estimates of the parameter in (2.5). This parameter speaks to the effectiveness of the programs, *as implemented*. Columns (5) and (8) display estimates of the parameters in (2.6) and (2.8). The former speaks to the effectiveness of the programs relative to the counterfactual of *staying at home*. The latter speaks to the effectiveness of the programs relative to *attending an alternative preschool*. All of these three estimates account for program attrition and control for a set of background variables.[38]

[38]See Appendix H for our methodology to account for attrition and Appendix B for our procedure for selecting controls.

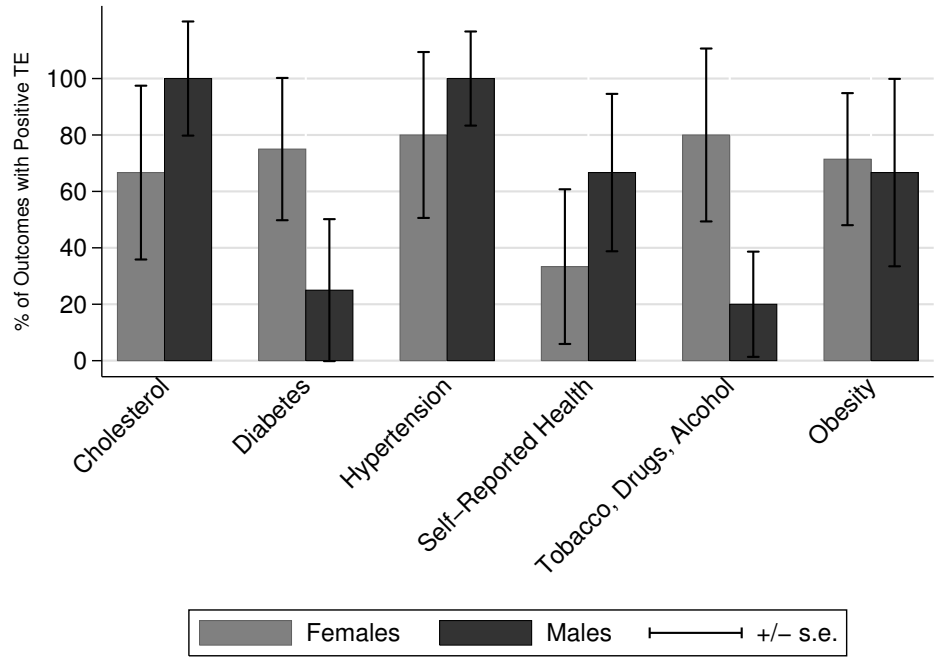
Figure 2.5: Positively Impacted Outcomes by Category, ABC and CARE



Note: For each outcome category, we compare the mean of the subjects who received center-based childcare in ABC and center-based childcare and family education in CARE to the mean of the subjects in the control group in both programs and count the number of positive comparisons.

Column (2) shows that the program caused substantial gains in a variety of economically relevant short- and long-term outcomes, as implemented. First, the programs have an effect on IQ that goes beyond the effects of many early childhood education programs, which usually fade out after a year of elementary school (Hojman, 2015; Elango et al., 2016). To put this effect in perspective, note that IQ and achievement tests scores are standardized to a nationally representative population with a standard deviation of 15 points. The effect of the programs we study amounts to more than 1/2 of a standard deviation. The largest effect of Head Start, for example, happens before elementary school and amounts to half of the effect of ABC and CARE at age 12, the latter effect being measured after elementary school (Elango et al., 2016). The programs also have a substantial effect on achievement, which not only measures cognition but mathematics and reading knowledge.

Figure 2.6: Positively Impacted Health Outcomes, ABC and CARE



Note: For each outcome category, we compare the mean of the subjects who received center-based childcare in ABC and center-based childcare and family education in CARE to the mean of the subjects in the control group in both programs and count the number of positive comparisons.

The effects that the programs have on years of education and employment at age 30 are sizable—column (2). The former increases by 8 percentage points and the latter by 1.7 years. Although marginally not significant, the programs also: (i) increase labor income; and (ii) reduce the dependence on public-transfer income. When we group the education and employment outcomes across ages 21 and 30 all of them display a positive treatment effect (see Figure 2.5).

When fixing the counterfactuals, a clear pattern emerges: females benefit much more from the programs relative to staying at home compared to how they benefit from the programs relative to attending alternative preschools. The differences are substantial: more than 30 percentage points in employment, almost 4 years of education, a decrease of around $\$3,000$ 2014 USD in public transfer income.

Table 2.8: Treatment Effects on Selected Outcomes, Females

Variable	Age	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Std. IQ Test	12	8.688	7.857	6.850	9.960	6.441	9.120	6.952	8.429
		(0.000)	**(0.013)**	**(0.026)**	**(0.000)**	**(0.039)**	**(0.000)**	**(0.039)**	**(0.013)**
Education Years	30	2.143	1.695	4.025	2.984	3.918	1.567	1.155	1.409
		(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.013)**	**(0.066)**	**(0.026)**
Transfer Income	30	-2,672	-1,560	-3,053	-2,783	-3,213	-2,269	-1,169	-2,620
		(0.026)	(0.118)	**(0.013)**	**(0.092)**	**(0.013)**	(0.105)	(0.224)	(0.132)
Employed	30	0.131	0.080	0.333	0.363	0.340	0.056	0.003	0.070
		(0.092)	(0.171)	**(0.053)**	**(0.092)**	**(0.053)**	(0.276)	(0.447)	(0.263)
Years Incarcerated	30	-0.024	-0.025				-0.037	-0.032	-0.038
		(0.053)	**(0.066)**				**(0.053)**	**(0.053)**	**(0.066)**
Diabetes	Mid-30s	-0.071	-0.032				-0.091	-0.039	-0.095
		(0.066)	(0.171)				**(0.066)**	(0.171)	**(0.039)**

Note: This table displays the treatment effects for females, pooling ABC and CARE. Column (1): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to. Column (2): adjusts the estimates in (1) for attrition and controls for a set of covariates (see Appendix B). Column (3): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, restricting the latter to subjects who did not receive preschool alternatives. Column (4) adjusts the estimates in (3) for attrition and controls for a set of covariates. Column (5): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, placing a relatively high weight on the subjects who are likely not to be enrolled in alternative preschools. Column (6): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, restricting the latter to subjects who received preschool alternatives. Column (7) adjusts the estimates in (6) for attrition and controls for a set of covariates. Column (8): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, placing a relatively high weight on the children who are likely to be enrolled in alternative preschools. The results in bold are significant at the 10% level in a single-sided, non-parametric, bootstrapped test.

When comparing the parameter estimates in columns (2), (5), and (8) we see that for females: the effectiveness of the program *as implemented* is lower than the effectiveness of the program relative to *staying at home*, while it is greater relative to attending alternative preschools. In an exercise analogous to that of Figure 2.4 we find that the programs relative to staying at home cause 84% (55%) positive (and significant outcomes). The analogous number relative to attending alternative preschool is 79% (33%). These results are relevant for the calculation of the cost-benefit ratio: they order the relative magnitude of the estimates we provide, depending on the counterfactual comparison.

Table 2.9: Treatment Effects on Selected Outcomes, Males

Variable	Age	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Education Years	30	0.525	0.708	0.857	1.302	0.791	0.385	0.540	0.347
		(0.079)	**(0.079)**	(0.118)	**(0.092)**	(0.171)	(0.171)	(0.132)	(0.276)
Labor Income	30	19,810	24,902	17,909	21,069	24,012	20,065	28,483	21,170
		(0.079)	(0.171)	(0.132)	(0.263)	(0.105)	**(0.066)**	(0.132)	(0.158)
Employed	30	0.119	0.179	-0.029	-0.050	0.041	0.176	0.245	0.262
		(0.079)	**(0.039)**	(0.487)	(0.579)	(0.355)	**(0.053)**	**(0.013)**	**(0.000)**
Misdemeanors	Mid-30s	-0.501	-0.239	-0.251	0.085	-0.040	-0.665	-0.343	-0.512
		(0.132)	(0.316)	(0.408)	(0.500)	(0.395)	(0.105)	(0.224)	(0.118)
Diastolic Pressure	Mid-30s	-10.854	-19.895	-8.640	-12.199	-8.150	-14.240	-22.740	-21.851
		(0.000)	**(0.000)**	**(0.013)**	**(0.079)**	**(0.026)**	**(0.013)**	**(0.000)**	**(0.000)**
Vit D Deficiency	Mid-30s	-0.245	-0.185	-0.480	-0.216	-0.485	-0.172	-0.145	-0.189
		(0.066)	(0.132)	**(0.000)**	(0.158)	**(0.000)**	(0.158)	(0.263)	(0.158)
Drug user	Mid-30s	-0.333	-0.432	-0.500	-0.788	-0.555	-0.233	-0.326	-0.330
		(0.026)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	(0.118)	**(0.053)**	**(0.039)**

Note: This table displays the treatment effects for females, pooling ABC and CARE. Column (1): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned as control. Column (2): adjusts the estimates in (1) for attrition and controls for a set of covariates (see Appendix B). Column (3): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, restricting the latter to subjects who did not receive preschool alternatives. Column (4) adjusts the estimates in (3) for attrition and controls for a set of covariates. Column (5): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, placing a relatively high weight on the subjects who are likely not to be enrolled in alternative preschools. Column (6): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, restricting the latter to subjects who received preschool alternatives. Column (7) adjusts the estimates in (6) for attrition and controls for a set of covariates. Column (8): mean difference between the groups randomly assigned to receive center-based childcare and the groups randomly assigned not to, placing a relatively high weight on the subjects who are likely to be enrolled in alternative preschools. The results in bold are significant at the 10% level in a single-sided, non-parametric, bootstrapped test.

Table 2.9 presents the results for males. The programs *as implemented* had statistically and economically significant effects on high-school graduation, employment, prediabetes, and diastolic pressure. The effect on labor income amounts to nearly $\$20,000$ 2014 USD.

When fixing the counterfactual comparison, a clear pattern, as in the case of females, is not evident. When complementing the information with counts of positive (and significant) treatment effects, a pattern does emerge: relative to the programs, males benefit more from *staying at home* than *attending alternative preschools*. Relative to staying at home, ABC

and CARE have positive (and significant) treatment effects for 47% (15%) of outcomes. The analogous number relative to attending alternative preschool is 79% (29%). The magnitudes of the results in the cost-benefit analysis are consistent with these results.

2.4.2 Cost-benefit Analysis

Table 2.10 summarizes the cost-benefit analysis of the programs without accounting for control substitution. All the money figures are in 2014 USD and are discounted to each child's birth age, unless otherwise specified.

Pooling males and females, the results indicate that the program is socially efficient: the baseline estimates for the internal rate of return and the benefit-to-cost ratio are 13% and 4.35. The program generates a benefit of 4.35 for every dollar spent on it. This is of particular importance because ABC and CARE were much more expensive than other early childhood education programs like Perry or Head Start (Elango et al., 2016)—the treatment involved more services over a longer time period.

The internal rate of return and the benefit-to-cost ratio are robust to sensitivity exercises. First, we remove the component due to parental income. In practice, ABC and CARE had a childcare subsidy component because it allowed the mothers to work causing additional parental income. This component amounts to $\$115,026$. Even after removing this component, the internal rate of return and benefit-to-cost ratio indicate social efficiency of the program and remain statistically significant.

Parental income and crime are the components for which the internal rate of return and the benefit-to-cost ratio are the most sensitive.[39] The reason for the sensitivity to parental

[39]We do not account for treatment effects on parental income beyond age 15 because some children report to move out of their households as early as this age. This makes ambiguous the effects that parental income could have on the subjects after age 15.

income is that the amount is substantial and it is not heavily discounted because it accumulates during the first 15 years of the children's life. Although crime is subject to more discounting, the amount due to crime savings is large so removing it diminishes both the internal rate of return and the benefit-to-cost ratio.

The estimates are robust to individually removing the rest of the components, and in most cases remain statistically significant. This happens for one of either two reasons: (i) the effects are substantial but they are heavily discounted because they happen later in life—e.g. labor income; or (ii) the effects happen early in life but are not as substantial—as in the amount that the control-group parents pay for their children to attend alternative preschools.

Next, we analyze the estimates when splitting the sample by males and females. Some of the estimates lose significance due to the reduction of observations after splitting the sample by gender. The point estimates remain robust across the sensitivity analysis.

For females, we observe consistency except for the case where we remove the parental income component. We can observe that the female sample is the main driver of parental income when comparing its net present value between females and males.

For males, the estimates are robust, similar to the female samples. An exception occurs when we remove the component corresponding to quality-adjusted life years. Although the cost-to-benefit ratio remains virtually unchanged, the internal rate of return is negative. In this case, the internal rate of return is actually uninformative: it is negative due to the fact that, when excluding the quality-adjusted life years, the net-benefit streams cross from negative to positive generating multiple roots. The interpretation of the internal rate of return when the net-benefit streams cross in unclear (Arrow and Levhari, 1969).[40]

[40]This reinforces the importance of considering the quality-life improvement due to better health conditions.

Table 2.10: Cost-benefit Analysis of ABC and CARE, Summary

	Females			Males			Pooled		
Removed Component	NPV	IRR	B/C	NPV	IRR	B/C	NPV	IRR	B/C
None	135,214	0.10	2.44	801,521	**0.15**	**10.78**	378,318	**0.13**	**5.73**
		(0.07)	(1.09)		(0.05)	(4.79)		(0.04)	(2.12)
Parental Income	112,402	0.04	1.05	91,722	**0.12**	**9.72**	115,026	**0.08**	**4.43**
		(0.02)	(0.97)		(0.04)	(4.79)		(0.02)	(2.00)
Subject QALY	59,203	**0.10**	2.36	213,866	**0.14**	**9.53**	127,014	**0.12**	**4.71**
		(0.08)	(1.02)		(0.06)	(4.42)		(0.05)	(2.09)
Subject Labor Income	39,587	0.09	1.97	191,868	**0.14**	**8.67**	112,510	**0.12**	**4.38**
		(0.08)	(0.79)		(0.05)	(3.86)		(0.05)	(1.88)
Subject Transfer Income	5,151	0.10	2.43	-908	**0.15**	**10.81**	-7,381	**0.13**	**5.76**
		(0.07)	(1.10)		(0.05)	(4.82)		(0.04)	(2.13)
Medical Expenditures	-51,391	0.10	2.52	-72,912	**0.15**	**11.15**	-66,496	**0.14**	**6.00**
		(0.08)	(1.13)		(0.04)	(4.78)		(0.04)	(2.13)
Control Contamination	16,725	0.08	2.25	13,543	**0.14**	**10.63**	13,879	**0.12**	**5.56**
		(0.05)	(1.09)		(0.05)	(4.79)		(0.04)	(2.12)
Education Costs	-41,792	**0.11**	**2.81**	12,234	**0.14**	**10.72**	-16,551	**0.13**	**5.86**
		(0.07)	(1.09)		(0.05)	(4.82)		(0.04)	(2.12)
Crime Costs	89,735	0.08	1.63	446,517	**0.10**	4.24	194,724	**0.10**	**3.36**
		(0.07)	(0.99)		(0.05)	(3.22)		(0.04)	(1.29)
Deadweight Loss		0.15	**3.51**		**0.18**	**16.04**		**0.19**	**8.51**
		(0.13)	(1.64)		(0.06)	(7.25)		(0.06)	(3.18)
0% Discount Rate			4.99			**27.06**			**14.53**
			(3.54)			(13.59)			(5.86)
5% Discount Rate			1.72			**6.37**			**3.51**
			(0.64)			(2.66)			(1.23)

Note: This table presents the estimates of the net present value (NPV) for each component, and the internal rate of return (IRR) and the benefit-to-cost ratio (B/C) of ABC for different scenarios based on comparing the groups randomly assigned to receive center-based childcare and the groups randomly assigned as control in ABC and CARE. The first row represents the baseline estimates. The rest of the rows present estimates for scenarios in which we remove the NPV estimates of the component listed in the first column. The quantity listed in the NPV columns is the component we actually remove when computing the calculation in each row. All the money figures are in 2014 USD and are discounted to each child's birth, unless otherwise specified. For the B/C we use a discount rate of 3%, unless otherwise specified. We test the null hypotheses IRR = 3% and B/C = 1—we elect 3% because that is the discount rate we use. The results in bold are significant at the 10% level in a single-sided, non-parametric, bootstrapped test. We resample both the experimental and the auxiliary data sources.

Finally, we provide a cost-benefit analysis of the program when accounting for control substitution (Table 2.11). The first row shows the estimates without accounting for control substitution, i.e. the same as those of the first row in Table 2.10. The second and third rows present results for the two counterfactual comparisons we consider.

Before discussing the results, it is worth noting that the sample sizes for some of the cases make the IRR estimates very unstable. For example, the estimates for females compared to the counterfactual of staying at home are based on an initial sample of 5 observations in the control group, while the estimates for males are based on 7 observations in the control group.[41] In the specific case of females, the internal rate of return corresponding to the counterfactual of staying at home is based on crossing net-benefit streams. Similarly, we cannot obtain a real solution for the case of males.

The samples used to produce estimates relative to enrollment in alternatives preschools are larger than those previously mentioned and we are able to obtain real solutions for the IRR even after splitting by males and females. Despite these practical difficulties, the results are consistent with the treatment effects we show in Section 2.4.1. Compared to ABC and CARE, females benefit more than males from alternative preschools relative to staying at home. That is, the benefit-to-cost ratio of ABC and CARE relative to staying at home is high for females and low for males. Conversely, the benefit-to-cost ratio of ABC and CARE relative to alternative preschools is high for males and low for females. Given the outcomes that we are able to monetize have higher values for males than for females, the pooled results are more similar to the results for males than to the results for females. Regardless, any of the counterfactual comparisons we consider indicates that ABC and CARE are socially efficient.

[41] That is, we observe 5 females and 7 males in the control group who did not attend alternative preschools. Some of the forecasts could be based on even smaller samples due to missing values in some specific outcomes.

Table 2.11: Cost-benefit Analysis Accounting for Control Substitution, ABC and CARE

	Females		Males		Pooled	
Estimate	IRR	B/C	IRR	B/C	IRR	B/C
Baseline	0.10	2.30	**0.15**	**7.88**	**0.13**	**4.35**
	(0.08)	(1.15)	(0.11)	(4.31)	(0.09)	(2.00)
Relative to Staying at Home	-0.14	**4.97**	0.02	0.55	**0.09**	**3.78**
	(0.11)	(1.58)	(0.07)	(2.25)	(0.03)	(1.68)
Relative to Alternative Preschools	0.08	1.58	**0.20**	**12.24**	**0.12**	**4.34**
	(0.08)	(0.93)	(0.13)	(4.16)	(0.07)	(2.14)

Note: This table displays estimates of the internal rate of return (IRR) and the benefit-to-cost ratio (B/C) for ABC and CARE for three cases. Not accounting for control substitution (baseline); comparing ABC and CARE to staying at home (relative to staying at home); and comparing ABC to alternative preschools (relative to alternative preschools). For the B/C we use a discount rate of 3%. We test the null hypotheses IRR = 3% and B/C = 1—we elect 3% because that is the discount rate we use. The results in bold are significant at the 10% level in a single-sided, non-parametric, bootstrapped test. We resample both the experimental and the auxiliary data sources.

2.5 Final Comments

The evidence from policies related to early childhood education is still scarce despite its importance in the public debate. We provide a thorough evaluation of two randomized controlled trials: the Carolina Abecedarian Project (ABC) and the Carolina Approach to Responsive Education (CARE).

As programs providing high-quality center-based childcare, ABC and CARE have positive effects on a variety of outcomes measuring human development throughout childhood to adulthood—including cognition, socio-emotional skills, criminal activity, and adulthood health. This translates into statistically and economically significant measures of social efficiency, like the benefit-to-cost ratio and the internal rate of return, which we calculate accounting for complications that arise when evaluating social programs and considering life-cycle gains.

When adequately assessed, early childhood education programs enhance human development in that they provide a vehicle to promote social mobility. An adequate assessment requires: (i) comparing the program with respect to a well-defined counterfactual—e.g. other programs or staying at home; and (ii) monetizing the life-cycle gains, which goes beyond back-of-the-envelope calculations based on short-term gains.

2.6 Acknowledgments

This research was supported in part by the American Bar Foundation; the Pritzker Children's Initiative, the Buffett Early Childhood Fund, NIH grants NICHD R37HD065072, NICHD R01HD54702, and NIA R24AG048081, an anonymous funder, Successful Pathways from School to Work, an initiative of the University of Chicago's Committee on Education funded by the Hymen Milgrom Supporting Organization, and the Human Capital and Economic Opportunity Global Working Group, an initiative of the Center for the Economics of Human Development, affiliated with the Becker Friedman Institute for Research in Economics, and funded by the Institute for New Economic Thinking. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders or the official views of the National Institutes of Health. Collaboration with Yu Kyung Koh, Sylvi Kuperman, Stefano Mosso, Rodrigo Pinto, and Anna Ziff on related work has strengthened the analysis in this paper. For helpful comments, we thank Stéphane Bonhomme, Flávio Cunha, Steven Durlauf, Azeem Shaikh, Matthew Tauzer, and Ed Vytlacil. For information on the implementation of the Carolina Abecedarian Project and assistance in data acquisition, we thank Peg Burchinal, Carrie Bynum, Frances Campbell, and Elizabeth Gunn. For information on childcare in North Carolina, we thank Richard Clifford and Sue Russell.

Chapter 3

Early Childhood Education

This chapter is coauthored with Sneha Elango, Jorge Luis García and James J. Heckman.

3.1 Introduction

Recent research demonstrates that the effects of adverse early childhood environments persist over a lifetime (Knudsen et al., 2006). Substantial gaps between the environments of advantaged children and those of disadvantaged children raise serious concerns about the life prospects of disadvantaged children and the state of social mobility in America.[1]

The proliferation of single-parent households—especially households where children have never had a father present—is a major contributor to the growth in inequality in childhood environments.[2] In the US, single-parenthood is strongly correlated with child poverty. As a group, the children of single parents are less likely to succeed in life than children from stable two-parent households.[3] This evidence and the evidence that gaps in advantage are growing across generations[4] has prompted growing interest in improving the early-life opportunities of disadvantaged children.[5]

Concerns about the quality of childhood environments are fueled by growth in the labor force participation of women with children.[6] This growth raises concerns about the supply of childcare and its quality. Disadvantaged parents often lack access to high-quality childcare and single-parent families are especially vulnerable.[7] The percentage of children who grow up in poverty has increased from 16% in 2000 to 21% in 2013.[8]

These dual concerns have stimulated interest in public provision of early childhood education programs to ease the burden of childcare for working mothers and to enhance the opportunities available to disadvantaged children.

[1] McLanahan (2004); Duncan and Murnane (2011).
[2] McLanahan (2004); Heckman (2008).
[3] McLanahan and Percheski (2008).
[4] Putnam (2015).
[5] Office of the Mayor, New York City (2014).
[6] Calculations using the Current Population Survey indicate that, between 1960 to 2010, maternal labor market attachment increased from 41% to 65% for single mothers (with children) and 20% to 60% for married mothers. Most of these single mothers had children residing with them—in 1960, 91% of children in single parent families lived with their mothers; this fell slightly to 87% in 2010.
[7] Blau (2003).
[8] Rates of child poverty are calculated using the Current Population Survey. Poverty is defined as growing up in a household below the federal poverty line.

High-quality early childhood education programs enrich the learning and nurturing environments of disadvantaged children. An accumulating body of evidence shows the beneficial effects of these programs. They are much discussed among academics, mainstream media, and policymakers. The Obama administration has promoted programs like Head Start as vehicles of opportunity and social mobility and has called for increased federal investment in high-quality programs developed and administered by states (The White House, 2014a).

This paper organizes and synthesizes the evidence on a variety of early childhood programs. We consider the evidence on means-tested programs.[9] Eligibility for these programs is determined by a measure of childhood poverty (either family income or close surrogates for it). We also consider the evidence on universal preschool programs.[10]

We gather in one place the evidence on the programs with the most rigorous evaluations for which the reported results can be replicated. We also devote some attention to the evidence from programs with flawed or limited evaluations, but do not place much weight on it. We compare the treatments, treated populations, and treatment effects across a broad range of programs.

We go beyond the standard, often very limited, discussions of the benefits of early childhood education. We consider a richer collection of outcome measures, in addition to the scores on IQ or achievement tests that receive so much attention in the literature. We consider multiple outcomes across the life-cycle, e.g., physical and mental health, criminal activity, earnings, and social engagement. We assess the economic and social rates of return for programs that have the necessary data.

We do not rely exclusively on evidence from randomized control trials. We use credible causal evidence from a broad range of studies using different methodologies. The evidence we assemble shows agreement across studies: there is a strong case for high-quality early

[9] "Means-Tested" in this paper refers to programs with eligibility criteria based on income, socio-economic status, or other measures of disadvantage.

[10] Universal programs have age requirements for children but are not means-tested. However, many advocate universal programs with sliding fee schedules based on family income, which effectively make them means-tested.

childhood education for disadvantaged children. It improves the early-life environments of disadvantaged children, which in turn boost a variety of early-life skills and later-life achievements.

We address two distinct questions that are frequently conflated. The first is whether or not early childhood programs are effective. The second is whether or not these programs should be subsidized by governments.

The answer to the first question depends on the quality of the program being offered and the alternatives available and their costs. Any measure of effectiveness is a relative statement. The proper question is: effective relative to what? Affluent families have better alternatives and generally do not benefit from the public provision of early childhood education aimed at median or disadvantaged populations. In contrast, high-quality versions of such programs are consistently found to benefit disadvantaged children and have substantial economic and social rates of return.[11]

Failure to account for the quality of childcare alternatives and the quality of home environments leads analysts to make misleading statements about program effectiveness. A recent example is the Head Start Impact Study (HSIS).[12] Analyses that fail to account for the childcare alternatives available to control participants understate the effects of Head Start. Analyses that account for these alternatives show that Head Start actually has moderate to strong effects on measures of cognitive and non-cognitive skills[13] compared to home care, but not necessarily when compared with other quality center-based childcare.

The answer to the second question is that the evidence in hand supports public subsidy of high-quality programs targeted to disadvantaged populations. At current quality levels and costs, their social benefits exceeds their social costs. There is little direct evidence on the effectiveness of the programs we study on the children of affluent families. This paper

[11]This conclusion is consistent with previous studies that argue that disadvantaged children greatly benefit from early childhood education. See, e.g., Blau and Currie (2006b), Duncan and Magnuson (2013b), and Yoshikawa et al. (2013b). We differ from these studies because we consider evidence from a broader range of studies using diverse but competent evaluation methodologies.

[12]Puma et al. (2012).

[13]Feller et al. (2016); Kline and Walters (2014); Zhai et al. (2014).

does not address the general question of what the optimal provision of childcare should be for persons in different economic strata. The answer to this question would take us too far afield.

The economic case for universal early childhood programs is weak.[14] The case often made for them is political in nature. Universality is sometimes sought to avoid stigma and to promote inclusion. The costs of offering such programs are diminished because, at the levels of quality usually proposed, the affluent are much less likely to use them.[15] The programs discussed in this paper are less attractive to them because they have better alternatives.

Table 3.1 summarizes the programs we discuss and their basic features. We present detailed descriptions of these programs in Sections 3.3–3.5 and Appendices A and B. Section 3.3 discusses the evidence from four experimental evaluations of demonstration programs: (i) the Perry Preschool Project (PPP); (ii) the Carolina Abecedarian Project (ABC); (iii) the Infant Health and Development Program (IHDP); and (iv) the Early Training Project (ETP). Instead of just reporting estimates from the literature, or doing a meta-analysis, we conduct a primary analysis of each program using a standardized format. We could not discuss the Chicago Parent-Child Program (Reynolds et al., 2011) in our analysis because we do not have access to the most updated and complete data for this program on which claims about its effectiveness are based . The PI has not cooperated to help us replicate its reported findings. Our access to data for the Nurse Family Partnership (NFP; Olds, 2006), is similarly restricted.

We consider the evidence on Head Start in Section 3.4. Eligibility for it is means-tested primarily on the basis of family income. Centers are free to pick their curricula and there

[14]Universal programs are defined as programs available to all children in a geographical area with only age as an eligibility criteria. Because they are voluntary, participation in universal programs is far from universal. For example, the take-up of the two major universal state programs in Georgia and Oklahoma for the years they are studied is 59% and 74%, respectively (Cascio and Schanzenbach, 2013). Within these programs, 65% and 66% of participating children were low-income as measured by eligibility for free or reduced price lunch, which is offered to children whose families are at or below 185% of the federal poverty line. We discuss preschool take-up by socio-economic status further in Section 3.5.

[15]Program costs would be diminished further if the affluent who used them were charged user fees, as some have proposed (Heckman, 2008).

is a lot of variety in the programs offered. We also discuss the evidence from a recently evaluated means-tested statewide program that shares some features in common with Head Start.[16]

The evidence on the benefits of universal programs discussed in Section 3.5 comes from: (i) national programs in Canada and Norway; (ii) state programs in Oklahoma and Georgia; and (iii) a recent universal program in Boston. Section 3.6 discusses non-experimental evidence on the importance of quality environments in promoting child development. We summarize our findings in Section 3.7.

The goal of this paper is to distill general lessons from the literature that can guide policy and not to endorse or attack any particular program. The literature is often marred by a "treatment effect" mentality that sees evaluation research as an up or down statement about whether a particular program "works" and not why it works or does not work. Our approach is to understand the mechanisms underlying successful early childhood education programs with an eye toward designing future approaches that improve on current practice. With this goal in mind, we next present a framework for interpreting the evidence within a general model of human development.

3.2 A Framework for Interpreting the Evidence

Before turning to our review of the literature, we present the guiding principles of this essay. We first discuss a dynamic model of skill formation based on Cunha and Heckman (2007, 2009). It provides a framework for understanding the effectiveness of early interventions for disadvantaged children. We next consider arguments for public provisions of interventions. We then discuss how the availability of alternative childcare options affects the interpretation of the evidence from interventions.

[16]The Tennessee Pre-Kindergarten Program (Lipsey et al., 2015).

Table 3.1: Comparing Demonstration Programs, Head Start, and Universal Preschool Programs

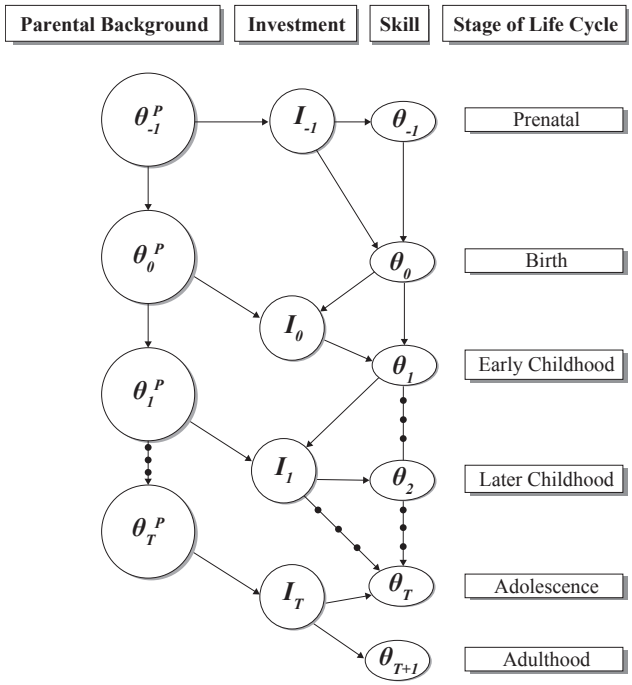
		Eligibility				Content				Sample Characteristics			Measures Available									
		Means-tested	High Disadvantage	Low Income	Criteria Narrowly Defined	Homogeneous Treatment	Medical Services	Home Visiting	Parent Involvement	Randomized Control Trial	Small Sample	Control Contamination	Age of Follow-ups	IQ	Achievement	Non-Cognitive	Parenting Skills	Subject Employment	Educational Attainment	Use of Public Transfers	Crime	Health
Demonstration Programs	ABC	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	34	✓	✓	✓	✓	✓	✓	✓	✓	✓
	PPP	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	-	40	✓	✓	✓	-	✓	✓	✓	✓	-
	ETP	✓	✓	✓	✓	-	-	✓	✓	✓	✓	-	20	✓	✓	-	-	-	✓	-	-	-
	IHDP	- [a]	-	-	-	✓	✓	✓	✓	✓	-	✓	18	✓	✓	✓	✓	-	✓	-	-	-
Head Start	HSIS	✓	-	✓	-	- [b]	✓	✓	✓	✓	-	✓	8	✓	✓	✓	-	-	-	-	-	-
	NLSY79/CNLSY	✓	-	✓	-	-	✓	✓	✓	-	-	✓	21	-	✓	✓	-	✓	✓	✓	✓	✓
Universal Programs	State Pre-K: OK	-	-	-	-	-	-	-	-	-	-	✓[c]	9	-	✓	-	-	-	-	-	-	-
	State Pre-K: GA	-	-	-	-	-	-	-	-	-	-	✓[c]	9	-	✓	-	✓	-	-	-	-	-
	Local Pre-K: Boston	-	-	-	-	-	-	-	-	-	-	✓[c]	6	-	✓	✓	-	-	-	-	-	-
	Reform in Norway	-	-	-	-	-	-	-	-	-	-	-	33	-	✓	-	✓	✓	✓	-	-	-
Other Programs	TN-VPK	✓	-	✓	-	-	-	-	-	✓	-	✓[d]	6	-	✓	✓	-	-	-	-	-	-

Note: This table compares the programs from which we draw evidence. **ABC**: Carolina Abecedarian Project. **PPP**: Perry Preschool Project. **ETP**: Early Training Project. **IHDP**: Infant Health and Development Program. **HSIS**: Head Start Impact Study. **TN-VPK**: Tennessee Voluntary Prekindergarten Program. **Boston**: Boston Public School Prekindergarten Program. "High Disadvantage" refers to inclusion of home environment and other family characteristics in the eligibility criteria. "Criteria Narrowly Defined" indicates that the program serves a population that is narrowly defined in terms of eligibility on the basis of socio-economic status or race. While Head Start serves predominately low-income children, the populations served vary greatly across sites in other important characteristics. "Homogeneous Treatment" refers to approximately equivalent quality across sites or cohorts. [a] IHDP limited participation to low birthweight, premature children ($\leq 2,500$ grams, ≤ 37 weeks) who lived at most 45 minutes away from treatment centers. [b] Although there are curricular guidelines and performance standards for Head Start, individual centers have flexibility in curriculum implementation and offer different services that are intended to meet the needs of the local population. Thus, we consider Head Start to have heterogeneous treatment, though there are similarities in treatment. Own calculations with HSIS data indicate that 30% of HSIS centers use a version of the *HighScope* curriculum, which was developed in the Perry Preschool Project. "Control Contamination" refers to the use by control children of other programs. There is some information on the nature of control contamination for almost all of the programs. [c] These programs are not randomized control trials. There is evidence a substantive part of the comparison groups in Boston and Oklahoma had access to center-based care. We assume that this can be extrapolated for the case of Georgia, where the information is less clear. [d] There is not much known about control contamination in TN-VPK; however, control children were not prohibited from enrolling in other programs. "Sample Characteristics" describe the features of the study design and data that impact evaluation. "Measures Available" describes the data available from our cited studies.

3.2.1 The Formation of Skills Over the Life-cycle

Cunha and Heckman (2007, 2009) develop a model of the evolution of skills over the life-cycle. The central ingredient of this model is the technology of skill formation, graphically represented in Figure 3.1. At life cycle stage t, parental skills ($\boldsymbol{\theta}_t^{\boldsymbol{P}}$), investment ($\boldsymbol{I_t}$), and child skills ($\boldsymbol{\theta_t}$) determine the skills in the next period $t+1$ ($\boldsymbol{\theta_{t+1}}$).[17]

Figure 3.1: Graphical Representation of the Technology of Skill Formation



Note: This figure illustrates the technology of skill formation, where links in the technology are represented by arrows. Dots represent periods that are not depicted in the diagram.

Parents affect their children in multiple ways. Parents with greater parenting skills ($\boldsymbol{\theta}_t^P$) create warm, supportive, fostering environments independent of their financial resources, the volume of time spent with children in direct instruction, or child development. Parents with greater financial and time resources can invest more in goods (e.g., tuition for pre-K) and time (e.g., taking a child to the zoo) captured by vector $\boldsymbol{I}_t$. Whether they choose to do so depends in part on their preferences.[18]

[17] $t = -1$ corresponds to the prenatal years.

[18] See, e.g., the review of the literature on parental preferences for child outcomes in Heckman and Mosso

Income is often used as a measure of child poverty, but it is a very crude one. An affluent but indifferent parent can provide an impoverished early childhood environment. Financially strapped families can nonetheless provide strong family environments through their attachment, warmth, and investment in time and caring. Public programs attempt to bolster both $\boldsymbol{I}_t$ and $\boldsymbol{\theta}_t^P$ and also to provide information to parents. While this paper focuses on "means-tested" programs, readers should recognize the inadequacy of equating childhood poverty with poverty in money income.[19]

The process of skill formation is dynamic and builds on itself. In the technology of skill formation, current stocks of skills help create future stocks of skills over the life-cycle, and future skills have intergenerational impacts. These dynamic relationships make early life an important period because it lays the foundation for building skills later in life. The following points are established in the recent literature.

1. *Skills are multiple.* Individuals have many life-relevant skills beyond the cognitive skills measured by IQ and achievement tests. These additional skills are variously referred to as non-cognitive skills or character skills. They also include health and mental health. They are important predictors of successful lives. These skills are important to different degrees in different life tasks. Early education programs promote these skills. In assessing the success or failure of any intervention, a full inventory of the skills affected is an essential part of any reliable evaluation of it.[20]

2. *Skills are self-productive and complement each other.* Between any two periods in the life of a child, t and $t+1$, a child's stock of skills builds on itself ("skills beget skills"). Skills are not only self-productive but also promote the production of other skills. Skills are said to complement each other in period t when together they promote skills in period $t+1$ more than each skill alone. Cognitive skills, non-cognitive skills, and health

(2014b).

[19]See Mayer (1997) and Heckman and Mosso (2014b).

[20]Heckman and Kautz (2012b, 2014).

in period t complement each other and produce cognitive skills, non-cognitive skills, and health in period $t+1$.[21]

3. *Skills complement investment.* By fostering early-life skills, early childhood education establishes a foundation which facilitates the accumulation of skills later in life.[22] Early childhood education promotes life-cycle skill development by increasing the stock of future skills that promote the productivity of future investment. This feature of life-cycle investment is called *dynamic complementarity.* Under conditions confirmed empirically in Cunha et al. (2010b), it is more productive to invest in disadvantaged children early in life than to remediate disadvantage later in life. This arises from the complementarity between later-life skills (acquired by early-life investment) and later-life investments. Enriched, early-life investment helps disadvantaged children capture many of the same benefits of later-life investment that are experienced by their more advantaged peers. The flip side of dynamic complementarity is that it is harder to remediate early disadvantage at older ages. Investment at later ages in adolescents lacking a strong early skill base is often much less productive than investment at early ages.[23]

These features of the technology of skill formation help to explain why supplementing parenting skills and the quality of investment offered to disadvantaged young children are socially fair and economically efficient strategies.[24]

3.2.2 Arguments for Subsidizing Early Childhood Education Programs

Many arguments have been made for subsidizing early childhood programs for disadvantaged families. Heckman and Mosso (2014b) summarize the literature.

[21] See, e.g., Heckman and Mosso (2014b).
[22] Cunha and Heckman (2008b); Cunha et al. (2010b).
[23] See Heckman and Kautz (2014).
[24] Heckman and Mosso (2014b).

All of the arguments build on the evidence that early childhood environments have profound consequences on the lives of children, and affect the entire society through reduced crime, enhanced health, greater educational attainment, and greater social engagement. Adverse early childhood environments create externalities—effects on society as a whole—that parents (for whatever reason) do not act on or internalize. The exact reasons for deficits in early investment are debated. There are three classes of arguments.

Some point to *borrowing constraints* facing disadvantaged families that have become more pronounced in recent decades with declining real wages for less educated workers and that are exacerbated by rising tuition costs (see Caucutt and Lochner, 2012 and Duncan and Murnane, 2014). Under this argument, parents under-invest in children because their cost of investing is greater than the social cost of funds. With the growth in single-parent families and the need for women to work to support their families, time constraints on parents have also increased.

The evidence on the importance of borrowing constraints is hotly debated (see, e.g., Mayer, 1997 and Heckman and Mosso, 2014b). As previously noted, more than money is involved in creating nourishing, productive child environments. The evidence that cash transfers to disadvantaged families have important effects on child development is weak.

Other *information-based* arguments have been advanced that note the importance of family knowledge of best practice child rearing.[25] There is considerable evidence that disadvantaged parents lack the information required to be effective parents. Many programs (ETP, IHDP, PPP) are based on this premise and it is one reason for home visiting programs. It is a justification for using in-kind transfers of information and direct supplements to parenting, rather than simple cash transfers.

More controversial is the argument that *parents lack sufficient altruism/concern* for their children. This paternalistic argument has evident merit in the case of abusive parents, or parents who deny children access to opportunities that would give them options the parents

[25]See Cunha et al. (2013) and Cunha (2015) for recent evidence on this question.

do not wish them to exercise (e.g., high school education for Amish children).

This paper does not evaluate the merits of these separate arguments. But the evidence shows that in contemporary American society, disadvantaged children face adverse child rearing environments, and high-quality targeted in-kind policies that have been implemented are effective.

3.2.3 Two Policy Evaluation Questions

In evaluating program impacts on skill development, researchers must be careful in understanding what the evidence reveals. Families differ in terms of the quality of the early environments offered to their children. Researchers need to distinguish between two questions when evaluating program effectiveness. The first question is: *What is the causal effect of an early childhood education program relative to a particular childcare alternative, where one of these alternatives might be no treatment at all?* The second question is: *What is the causal effect of adding a program to the available choice set?*[26]

The first question addresses the effectiveness of a policy that offers a particular early education program compared to a particular alternative, e.g., home care. The second question addresses the effectiveness of *expanding* the choice set available to parents, i.e., adding one more alternative. Most of the evaluations we consider answer the second question, even though answers to it are often treated as answers to the first.[27]

These questions are often confused. In particular, estimating the causal effect of expanding the availability of choices—making a new program available—and interpreting such estimates as statements about the effectiveness of that program compared to no program at all, might suggest that a program is ineffective. If the control group of a study has access to alternatives that are good substitutes for the program being studied, and if the researcher erroneously assumes that the relevant alternative to the program being evaluated

[26]See Heckman and Vytlacil (2007).

[27]Heckman et al. (2000) discuss these problems under the rubric of "substitution bias." See also Heckman (1992).

is home childcare and not some higher quality alternative, then there would appear to be no causal effect of the program's availability—even though the program may be highly effective compared to home child care.[28]

This type of error is made in many evaluations of Head Start—particularly, in evaluations that use data from the Head Start Impact Study (HSIS). The control group in HSIS had access to treatment substitutes, which sometimes include other Head Start centers. Studies that ignore the availability of program substitutes find weak effects.[29] Studies that account for the substitutes available find moderate to strong effects of Head Start compared to no program at all on measures of cognitive skills and non-cognitive skills.[30]

We discuss this evidence in detail in Section 3.4 after discussing the evidence from demonstration programs. A discussion of these programs is relevant to our analysis of Head Start. The curricula of these programs are embedded in versions of the curricula used in Head Start centers, although they are funded at lower levels than in the original programs. Our evidence on demonstration programs offers indirect evidence on the possibilities for success of an enriched Head Start program.

3.3 Evidence from Demonstration Programs

This section analyzes the evidence from the demonstration programs listed in Table 3.1. We conduct a new primary analysis of the four programs listed there rather than just a meta-analysis of existing studies. We first present the common features of the demonstration programs we analyze and our criteria for selecting them. We then describe them in Subsection 3.3.2. We discuss common methodological issues that arise when analyzing these programs in Subsection 3.3.3. In Subsection 3.3.4 we present evidence on the short-term effects from these programs. We present evidence on long-term effects in Subsection 3.3.5. Subsection 3.3.6 relates the short-term findings to the long-term findings. Subsection 3.3.7

[28] See Heckman et al. (2000).
[29] Puma et al. (2012).
[30] Feller et al. (2016); Kline and Walters (2014) and Zhai et al. (2014).

discusses cost-benefit analyses for two major demonstration programs, PPP and ABC. Subsection 3.3.8 summarizes the discussion.

3.3.1 The Characteristics of the Demonstration Early Childhood Programs

The early childhood demonstration programs we consider are targeted social experiments designed to bolster various aspects of the early lives of disadvantaged children. Assignment to treatment is randomized, although non-compliance and attrition can compromise the inference from any randomization. These programs are all means-tested, though they have different eligibility criteria.

The evidence on demonstration programs is not always comparable across programs, because they differ in terms of data availability, eligibility, quality, duration of treatment, length of follow-up, and other characteristics. Careful analysis is required in making valid cross-program comparisons of program effects. We discuss program differences and identify common components. The demonstration programs considered here have the following common features:

1. *They are center-based.* This section focuses on four center-based programs: (i) the Perry Preschool Project (PPP); (ii) the Carolina Abecedarian Project (ABC); (iii) the Infant Health and Development Program (IHDP); and (iv) the Early Training Project (ETP).[31]

2. *They are means-tested.* The programs we consider are all means-tested, although they

[31] We do not consider three important programs outside of the US: the Mauritius Study, due to its excessive attrition by age 40 (58%) (Raine et al., 2010), the Turkey Early Enrichment Program, also due to its excessive attrition by age 26 (49%) (Kagitcibasi et al., 2009), and the Jamaica Study (Gertler et al., 2014), which focused primarily on nutrition and home visits. We do not consider the Nurse Family Partnership program because it focused mainly on prenatal care (Olds et al., 1986, 1994; Eckenrode et al., 2010; Heckman et al., 2014). Other programs in the US that we do not consider include the following: the Milwaukee Project, because data are unavailable (Page, 1972; Sommer and Sommer, 1983; Garber, 1988; Gilhousen et al., 1990); the Even Start Program (Ricciuti et al., 2004) and the Comprehensive Child Development Program (St. Pierre et al., 1999, 1997) because of lack of information on childcare alternatives.

use different eligibility criteria. The evidence on universal programs discussed in Section 3.5 shows that early childhood education is particularly effective for disadvantaged children.

3. *The programs considered collect measurements on multiple skills and outcomes over long periods of the life-cycle.* It is a common but mistaken practice to evaluate programs based on outcomes only measured at early ages. Uninformed analysts sometimes assume that programs are ineffective due to the fadeout in IQ in the short-term evaluations that ignore multiple capacities. We evaluate programs using a diverse set of long-term outcomes that matter for success in life, such as health, education, earnings, and participation in crime.

4. *We discuss, where necessary, the consequences of compromised randomization, attrition of participants from programs or from study samples, the availability of good substitutes in the control group, and other challenges in conducting evaluations.* Compromises of the initial randomization protocols occur when subjects assigned to treatment or control status in an experimental protocol switch their initially assigned status or leave the program or the follow-up surveys. Despite challenges in analyzing the data, we show that valid, policy-relevant information can be derived from these studies.

3.3.2 Overview of Programs Discussed in This Section

Table 3.2 presents an overview of the programs we study. We discuss their most prominent characteristics in the next few paragraphs and present a more detailed discussion in Appendix A. The oldest programs we study are ETP and PPP. They began in 1962 and continued until 1964 and 1967, respectively. ABC is also relatively old, beginning in 1972 and continuing until 1982. The most recent program is IHDP, implemented from 1985 to 1988. PPP and ABC have high-quality data with long-term follow-ups. IHDP and ETP only have follow-ups into young adulthood. ETP, PPP, and ABC shared a common goal of

preventing "mental retardation" and promoting school-readiness (Weikart, 1967; Gray et al., 1982b; Ramey et al., 1982; Zigler and Muenchow, 1994).[32]

The researchers who implemented ETP, PPP, and ABC also created the curricula for these programs. The staff adapted and improved them while they were being conducted (Heckman et al., 2015). All three curricula have elements in common: promotion of play-based and child-directed learning, emphasis on language development, and emphasis on developing non-cognitive and problem-solving skills. The curricula in IHDP was adapted from the curricula of both ABC and a spinoff program, the Carolina Approach to Responsive Education (CARE) (Gross et al., 1997).[33]

Of these studies, PPP and ABC presently have the longest follow-ups, with data up to ages 40 and 34, respectively. A follow-up through age 50 of Perry is being collected at the time of this writing. Both PPP and ETP served preschool-age children and had home visits with their parents. ABC served children from birth through preschool age. IHDP served children and had home visits from birth to age 3. ABC had two treatment phases, 0 to 5 and 5 to 8, and correspondingly two rounds of randomization. ABC was the most intensive program (8 hours per day starting from 1-3 months and continuing to age 8). There were no home visits in the first phase but parents were encouraged to visit the center. There were home visits in the second phase. We focus on the first phase (0-5) because there is little evidence of treatment effects from the second phase.[34] While ETP, PPP, and ABC served relatively narrowly targeted populations, IHDP was more inclusive and served a population that was far more heterogeneous in terms of race and socio-economic status, although all children served had low birth-weight.[35]

All four programs had relatively educated staffs with some experience in education and

[32]Note that the clinical understanding of mental retardation was once associated with disadvantages that hindered early life development Noll and Trent (2004).

[33]Appendix C provides further details about CARE.

[34]See Yi et al. (2015) and Campbell et al. (2014).

[35]García (2015) compares the IHDP sample with the cohort born in the same year (1985) in the US. The author finds that IHDP individuals are, on average, relatively disadvantaged. The author suggests that this is a consequence of the correlation between measures of disadvantage: maternal labor supply, household income, a father's presence at home, premature birth status, and low birth-weight.

high teacher-to-child ratios. They varied in the amount of time children spent in the center—PPP had 2 years of center-based treatment for 3 hours a day and weekly home visits; ETP had intensive summer school and weekly home visits during up to 3 years, but no year-round center care; and ABC included center-based care during all of early childhood from birth to school entry for up to 8 hours per day.

Like ABC, IHDP also began at birth. During the first year, the program provided weekly home visits. These visits became bi-monthly in the second and third years of treatment. IHDP provided center-based treatment for up to 9 hours a day for 50 weeks a year in the second and third years of the program. Both ABC and IHDP included medical components—most prominently regular physical check-ups for the treated children.

Table 3.2: Summary Table of Demonstration Programs

	PPP	ABC	IHDP	ETP
Program Overview[a]				
Years implemented	1962–1967	1972–1982	1985–1988	1962-1964
Site	Ypsilanti, Michigan	Chapel Hill, North Carolina (UNC)	8 sites selected after competitive review	Segregated black schools in Abbotsfield, Tennessee
# Cohorts	5	4	1	2
N (Treatment : Control)	123 (58 : 65)	111 (57 : 54)	985 (377 : 608)	88 (43 : 45)
Age of Entry	3–4	0	0	4–5
Duration	1–2 years	5 years	3 years	2–3 years
Treatment				
Home Visits[b] (per month)	4	0	4 (up to age); 1–2 (after age 1)	4
Center Care (weeks per year)	30	50	50	10
Center Care (hours per week)	12–15	45	20+	20
Parent Involvement	✓	-	✓	-
Nutrition	-	✓	✓	-
Diapers/Child Care Goods	-	✓	✓	-
Well-child Health Care	-	✓	✓	-
Ill-child Health Care	-	✓	✓	-
Counseling	-	✓	✓	-
Parenting Instruction	✓	-	✓	✓
Control[c]				
Home Visits (per month)	-	-	-	-
Center Care (weeks per year)	-	-	-	-
Center Care (hours per week)	-	-	-	-
Nutrition	-	✓(Formula up to 15 mo)	-	-
Diapers (no other health care goods)	-	✓(up to 15 mo)	-	-
Well-child Health Care	-	✓(Cohort 1, up to age 1)	✓	-
Ill-child Health Care	-	-	-	-
Counseling	-	-	-	-
Parenting Instruction	-	-	-	-

(Continue)

(Continuation)

	PPP	ABC	IHDP	ETP
Randomization Protocol				
Steps	1. Rank by initial IQ of child 2. Group evens and odds 3. Balance gender, SES, etc. 4. Randomize whole group	1. Match on HRI[d] 2. Adjust by gender, maternal IQ, siblings 3. Randomize pairs	1. Stratify on birthweight and site 2. Randomize	Simple randomization into 2 treatment and 1 control groups
Compromises	Enrolled siblings receive same assignment Working moms switched to control	2 extremely needy switched to treatment 4 refused random assignment 4 abandoned treatment 2 considered ineligible after randomization	17 families refused to participate of the study after assignment	N/A
Counterfactual	Stay at home or with friends or relatives (Few substitutes)	Stay at home or childcare Alternative programs available	Stay at home or childcare Alternative programs available	Stay at home or with friends or relatives (Few substitutes)
Program Eligibility[e]	Cultural Deprivation Scale < 11 Low IQ (< 85) African-American No physical handicap	HRI ≥ 11 Biologically healthy No signs of mental retardation	Live within 45 min from center Birth weight < 2500g Gestational age < 37 weeks No severe illnesses or neurological defects	Home environment: Education of parents Parent occupation semi- or unskilled African American Parent edu $\leq$ high school
Curriculum				
Adult-Child Ratio	1:5–1:6	1:3 (age 0–1); 1:4-5 (age 1–4); 1:5-6 (age 4–5)	1:3–1:4	1:4–1:6
Staff & Certifications				
Teachers	B.A.[g]	HS grads, mixedf	College grads	Teaching Assistants, college & Ph.D. students
Specialists	Special Ed. Teachers[g]	Physician, Nurse M.A.[f]	College grads[f] Clinical staff	Home visitors[f, g]
Language Development	✓	✓	✓	✓
Motor Development	-	✓	✓	-
Cognitive Development	✓	✓	✓	✓
Non-Cognitive Development	✓	✓	✓	✓
Task Orientation	-	✓	-	✓
High-Risk Behavior	-	✓	-	-
School Readiness	✓	✓	✓	✓

Source: All details and sources are extensively discussed in Appendix A. Notes: [a] In IHDP, an additional 105 twins were also followed in the study, but are not analyzed in the literature. These twins were assigned to the same treatment group as their siblings. For each site, the program lasted until the youngest child turned 36 months old, correcting for prematurity. [b] In PPP, home visits were intended to involve the mother in educating the child, increase her understanding of the educational process, and to extend the curriculum beyond the classes and into the homes. Monthly group meetings for parents were also available, but is not well documented. During IHDP home visits, families in treatment groups were given toys with instructions on how to play with their child with the toys. This was to extend the curriculum beyond the classroom. Home visits also sought to improve the parents ability to problem solve, cope with personal issues, and function as parents. In addition, parent groups were offered as a chance for parents to share information and concerns with each other, and to provide them with the opportunity to learn about child education and community resources. Surveys were conducted by college graduates. ETP had two treatment groups. In one group, parents received two 9-month training sessions; in the other, parents received one 9-month training session. During these training sessions, the objective of the intervention was made clear to mothers during visits to schools. Mothers were encouraged to engage in their children's learning, as well as to expand the experiential environment of the child (e.g. trips to the library). [c] Treatment group individuals received all these items as well. The control group of the first cohort of ABC received health check-ups for the first year, after which this practice was discontinued. [d] In ABC, the High Risk Index (HRI) was comprised of: "Absence of maternal relatives in the area"; "Siblings of school age one or more grades behind age-appropriate level or with equivalently low scores on school-administered achievement test"; "Payments received from welfare agencies within past 3 years"; "Record of father's work indicates unstable or unskilled and semiskilled labor"; "Record of mother's or father's IQ indicate scores of 90 or below"; "Record of sibling's IQ indicates scores of 90 or below"; "Relevant social agencies in the community indicate the family is in need of assistance"; "One or more members of the family has sought counseling or professional help in the past 3 years"; maternal and paternal educational levels; family income; father's presence. [e] In PPP, criteria for home environment included education of parents, occupational level of father, maternal employment, and household density. [f] Signifies that staff were specially trained for the program. [g] Signifies that staff were state certified.

PPP, ABC, and ETP are not strictly means-tested programs. They use varying measures of disadvantage roughly correlated with income, such as the quality of home environments as characterized by single parenthood, parental education, and housing density. Additionally, PPP and ETP were explicitly designed to serve African-American children.

IHDP differs from the other programs in its eligibility criteria. All participants were premature births ($\leq$ 37 weeks), low birth-weight ($\leq$ 2500 grams), and resided, at most, 45 minutes away from the location of the program. While the other demonstration programs served fairly narrowly defined disadvantaged populations (although the criteria used differ), IHDP served a population that was more heterogeneous in socio-economic status and race and only homogeneous in child birth-weight. However, because perinatal health is related to the socio-economic characteristics of the parents, IHDP subjects were disadvantaged compared to the general US population (García, 2015). Table 3.3 describes the baseline characteristics of the populations served by the four demonstration programs we study.[36]

[36]We describe only the control groups.

Table 3.3: Control Group Background Characteristics at Baseline, All Programs (Mean Outcomes)

	PPP		ABC		IHDP		ETP	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Black	100%	0%	97%	16%	53%	50%	100%	0%
IQ, Ages 2–4	79.02	6.44	90.42	11.46	88.00	20.16	87.29	11.88
Mother's Age	29.10	6.57	19.89	4.82	24.87	6.00	30.11	8.84
Mother's Years of Education	9.42	2.20	10.23	1.84	12.40	2.42	8.96	2.62
Mother Works	20%	40%	73%	45%	34%	47%	40%	49%
Father at Home	53%	50%	29%	46%	56%	50%	87%	34%
Father's Age	32.81	6.88	23.21	5.91	27.64	6.67	32.82	10.10
Father's Years of Education	8.60	2.40	10.95	1.76	13.16	2.89	9.59	2.75
Father Works	86%	35%	87%	34	51%	50	97%	17%
Household Income (2014 USD)	$\frac{1}{n}$	$\frac{1}{n}$	7,653	10,049	41,868	32,623	$\frac{1}{n}$	$\frac{1}{n}$
Siblings	4.28	2.59	0.64	1.10	1.02	1.17	3.59	2.21
Treatment	47%	50%	52%	50%	39%	49%	48%	50%

Source: Own calculations. Note: This table displays baseline characteristics of the control group of the demonstration programs we study. Mother and father's years of education are counted as the number of years of schooling completed by the mother and father, respectively, at the time of program entry. The number of siblings is reported at program entry. **PPP**: Child's IQ at age 3 is measured using the Stanford-Binet Intelligence Scale. **ABC**: Child's IQ at age 2 is measured using the Stanford-Binet Intelligence Scale. Mother's age is reported at the time of program entry. **IHDP**: Child's IQ at age 3 is measured using the Stanford-Binet Intelligence Scale. **ETP**: Child's IQ at age 4 is measured using the Stanford-Binet Intelligence Scale. Test scores are constructed to have a national mean of 100 and a standard deviation of 15. We only report characteristics of the control group, because for programs that started at birth (ABC and IHDP), we do not observe treatment baseline characteristics. Household income was not an eligibility criteria in any of the programs in this table. $\frac{1}{n}$ indicates this data was not available.

3.3.3 Possible Limitations in the Evidence from Demonstration Programs

Age of Programs

The programs we study are valuable for analyzing the effectiveness of early childhood education because long-term follow-ups of their participants are available. Though it is natural to question the relevance of older programs to current policy, we argue that the lessons from them are still highly relevant.

The basic principles of enhancing the investments in, and the environments of, disadvantaged children that were laid down fifty years ago remain intact. Objections to relying on evidence from early high-quality programs are made by analysts who think that the outcome

of an evaluation study should be an up or down assessment of *that* program, rather than a contribution to understanding the general principles from multiple programs that can guide the construction of future programs. The effectiveness of any particular program is presumably a lower bound on the effectiveness of new programs that build on and improve that program. Evidence for the success of a program should not be a call for slavish application of that program.

We make four additional points on the relevance of the evidence from older programs. First, all of the demonstration programs we analyze have school-readiness as a main goal. This goal is shared with most contemporary early education programs. Second, the success of some of these demonstration programs influenced the creation and design of the most important current early childhood education programs. ETP and PPP influenced the creation of Head Start (Zigler and Muenchow, 1994), and ABC motivated policymakers to consider programs that targeted even younger children and inspired the creation of Early Head Start (Schneider and McDonald, 2006). Third, and most important, as documented in Section 3.4.1, although demonstration programs were very high-quality for their time, they bear strong resemblance to current high-quality early childhood education programs in terms of their structure, staffing, and curricula. For example, a version of *HighScope* is the second most commonly used curriculum in Head Start, utilized by roughly 30% of Head Start centers.[37] Contemporary programs share other features with the programs we study, such as teacher-to-child ratios (Heckman et al., 2014). Finally, some of the programs studied have long-term follow-ups. Understanding the impacts of early childhood education on skill formation requires analysis of effects on adult outcomes. This research requirement necessitates analysis of older programs. Positive long-term outcomes are a strong indication of a well-designed program.

[37]Our own calculations using HSIS data.

Small Sample Sizes

Samples are often small. Several recent studies use exact small sample inference to estimate multiple treatment effects with precision, even when dividing samples by gender and accounting for the biases arising in testing multiple hypotheses ("cherry picking").[38] Application of small sample inference methods produces results that are often not substantively different from the results using bootstrap or standard asymptotic inference procedures (Heckman et al., 2010c; Campbell et al., 2014). The methodologies employed to analyze IHDP, PPP, and ABC are conservative.

Control Contamination

The extent to which the control group received center-based care varies across ETP, PPP, ABC, and IHDP. There was no control contamination in ETP or PPP because of a lack of center-based substitutes, whereas there was control contamination in ABC and IHDP which were launched after Head Start was founded. In ABC, the control group had access to non-center-based and center-based childcare, especially during ages 0–5 (García et al., 2015). This included high-quality care provided in churches and even care at one Head Start center. In IHDP, 39% of the children attended substitute programs, though their quality is unknown (García et al., 2014). None of the studies we discuss address the issue of control contamination, even though most of the control groups had access to high-quality alternatives. This practice makes conservative reported estimates of the effects of the programs (compared to the home alternative).

[38]See Romano et al. (2010). If a 10% significance level is used in a sample with 100 outcomes, and thus 100 null hypotheses of no treatment effects, roughly 10 would be "statistically significant" even if all null hypotheses are true, i.e., treatment had no effect on any outcome. Heckman et al. (2010c); Gertler et al. (2014); Campbell et al. (2014) and Heckman et al. (2014) use methods to correct for this multiplicity of hypotheses.

Attrition and Non-Response

PPP and ABC data are used for assessing long-term benefits because they have high-quality follow-ups. Follow-ups are available through age 40 in PPP and through age 34 in ABC. Attrition and non-response complicate the interpretation of the evidence. Reliable analyses adjust for these features of the data.

3.3.4 Effects on IQ, Achievement Test Scores, and Conscientiousness

Table 3.4 presents estimated treatment effects on early IQ, early and late achievement test scores, and early conscientiousness pooled over genders. Tables 3.5 and 3.6 display the same information by gender. We adjust all test statistics for the effects of multiple hypothesis testing using procedures applied in Heckman et al. (2010c). We base our interpretation on non-parametric, permutation-based, one-sided p-values to test if the programs had positive effects on the outcomes described. However, we also report results using two-sided tests. Effects are shown for two measures of cognition: IQ and achievement test scores. All effects are presented in units of standard deviations. In the case of IQ, we follow the convention and use standardized scores that normalize the population mean and standard deviations of 100 and 15, respectively. Also shown are effects on conscientiousness, a non-cognitive skill that is of interest due to its low correlation with cognition and high correlation with important later-life outcomes (Borghans et al., 2008; Heckman et al., 2014).

Table 3.4: Treatment Effects on Early-life Skills for Samples Pooled Across Gender

		Treatment Effect	Permutation, one-sided	Permutation, two-sided	Stepdown, one-sided	Stepdown, two-sided
Perry	IQ, Age 5	11.422	**0.000**	**0.000**	**0.000**	**0.000**
	IQ, Age 8	1.254	**0.080**	0.430	**0.080**	0.430
	Achievement Test Score, Ages 5–10	0.394	**0.000**	**0.000**	**0.010**	**0.010**
	Conscientiousness, Ages 4–7	0.273	**0.040**	**0.060**	**0.050**	**0.070**
	Achievement Test Score, Age 27	1.795	**0.020**	**0.070**	**0.080**	**0.060**
ABC	IQ, Age 5	6.398	**0.030**	**0.030**	**0.030**	**0.030**
	IQ, Age 8	4.500	**0.080**	**0.080**	0.180	0.180
	Achievement Test Score Ages 5–10	0.544	**0.010**	**0.010**	**0.020**	**0.020**
	Conscientiousness Ages 4–7	0.047	0.400	0.680	0.860	0.890
	Achievement Test Score, Age 21	0.422	**0.010**	**0.010**	0.120	0.120
IHDP	IQ, Age 3	8.475	**0.000**	**0.000**	**0.000**	**0.000**
	IQ, Age 8	-0.671	0.680	0.420	0.910	0.430
	Achievement Test Score, Ages 5–10	-0.012	0.570	0.840	0.830	0.870
	Conscientiousness, Ages 4–7	0.075	**0.060**	0.140	0.180	0.190
	Achievement Test Score, Age 18	0.108	0.470	0.950	0.730	0.930
ETP	IQ, Age 7	6.343	**0.020**	**0.080**	**0.050**	**0.050**
	IQ, Age, 8	5.743	0.100	0.240	0.150	0.200
	Achievement Test Score, Ages 5–10	0.534	0.380	0.820	0.510	0.800

Source: Own calculations. Note: Initial sample sizes are: PPP: 123; ABC: 122; IHDP: 985; ETP: 91. Non-parametric permutation $p-values$ account for compromised randomization, small sample size, and item non-response. See Heckman et al. (2010c) and Campbell et al. (2014, appendix) for details. Stepdown $p-value$ accounts for the same and for multiple hypotheses testing. All school-age and adult achievement and conscientiousness measures have mean 0 and standard deviation 1. All IQ measures have mean 100 and standard deviation 15 and they are standardized using the national population mean and standard deviation. For PPP, IHDP, and ETP at ages 5, 3, and 7 we use the Stanford-Binet IQ test. For ABC at 5 we use the Wechsler Preschool and Primary Scale of Intelligence. For PPP and ETP at age 8 we use the Stanford-Binet IQ test. At this same age, we use Wechsler Intelligence Scale for Children for ABC and IHDP. School Age Achievement is a factor measured through a factor of items at ages 5, 6, and 7. The items analyzed come from the California Achievement Test (ABC, PPP); Metropolitan Achievement Test (ETP); Peabody Individual Achievement Test (ABC); Woodcock-Johnson Test of Achievement (ABC, IHDP). School Age Conscientiousness is a factor constructed through a battery of items from various questionnaires: Achenbach Child Behavior Checklist (ABC); Classroom Behavior Inventory (ABC); Walker Problem Behavior Identification Checklist (ABC); Teacher rating (PPP, IHDP); Reputation test (PPP, IHDP). Adult achievement is measured by Adult Performance Level (PPP); WoodcockJohnson Test (ABC); Wechsler Adult Intelligence Scale (IHDP). Adult achievement and conscientiousness measures are not available in ETP.

Table 3.5: Treatment Effects on Early-life Skills for Females

		Treatment Effect	Permutation, one-sided	Permutation, two-sided	Stepdown, one-sided	Stepdown, two-sided
Perry	IQ, Age 5	12.666	**0.000**	**0.000**	**0.000**	**0.000**
	IQ, Age 8	4.240	0.410	0.900	0.700	0.940
	Achievement Test Score, Ages 5–10	0.564	0.180	0.400	0.300	0.390
	Conscientiousness, Ages, 4–7	0.515	0.380	0.850	0.610	0.860
	Achievement Test Score, Age 27	0.407	0.110	0.390	0.330	0.430
ABC	IQ, Age 5	3.051	**0.050**	**0.050**	**0.060**	**0.060**
	IQ, Age 8	4.573	0.110	0.150	0.360	0.360
	Achievement Test Score, Ages 5–10	0.822	0.260	0.280	0.410	0.410
	Conscientiousness, Ages 4–7	0.110	0.600	0.960	0.910	0.960
	Achievement Test Score, Age 21	0.737	0.240	0.600	0.790	0.840
IHDP	IQ Age 3	9.877	**0.000**	**0.000**	**0.000**	**0.000**
	IQ Age 8	-0.158	0.780	0.490	0.940	0.600
	Achievement Test Score Ages 5–10	-0.034	0.500	0.920	0.790	0.970
	Conscientiousness, Ages 4–7	0.089	0.240	0.440	0.500	0.530
	Achievement Test Score, Age 18	0.517	0.650	0.790	0.840	0.910
ETP	IQ, Age 7	8.611	0.120	0.140	0.180	0.180
	IQ, Age 8	9.056	0.290	0.540	0.440	0.550
	Achievement Test Score, Ages 5–10	0.448	0.810	0.350	0.980	0.270

Source: Own calculations. See notes in Table 3.4.

Table 3.6: Treatment Effects on Early-life Skills for Males

		Treatment Effect	Permutation, one-sided	Permutation, two-sided	Stepdown, one-sided	Stepdown, two-sided
Perry	IQ, Age 5	10.607	**0.000**	**0.000**	**0.010**	**0.010**
	IQ, Age 8	-0.721	**0.060**	0.250	0.150	0.190
	Achievement Test Score, Ages 5–10	0.269	**0.000**	**0.020**	**0.050**	**0.050**
	Conscientiousness, Ages4–7	0.087	**0.030**	**0.040**	**0.040**	**0.040**
	Achievement Test Score, Age 27	0.214	0.110	0.230	0.160	0.200
ABC	IQ, Age 5	9.962	0.530	0.540	0.890	0.890
	IQ, Age 8	4.174	0.410	0.410	0.760	0.760
	Achievement Test Score, Ages 5–10	0.277	**0.010**	**0.010**	**0.030**	**0.030**
	Conscientiousness, Ages 4–7	0.009	0.590	0.690	0.980	0.980
	Achievement Test Score, Age 21	0.095	**0.070**	**0.070**	0.120	0.120
IHDP	IQ, Age 3	6.988	**0.000**	**0.000**	**0.000**	**0.000**
	IQ, Age 8	-1.206	0.450	0.930	0.810	0.950
	Achievement Test Score Ages 5–10	0.012	0.720	0.650	0.900	0.740
	Conscientiousness, Ages4–7	0.065	**0.090**	0.170	0.250	0.270
	Achievement Test Score, Age18	-0.456	0.500	0.820	0.710	0.840
ETP	IQ, Age 7	4.111	0.100	0.200	0.160	0.170
	IQ, Age 8	2.333	0.140	0.210	0.260	0.280
	Achievement Test Score, Ages 5–10	-0.795	0.180	0.280	0.260	0.280

Source: Own calculations. See notes in Table 3.4.

All programs have positive effects on early measures of IQ. For both females and males in PPP, this effect is approximately 3/4 of a population standard deviation. The effects are also sizable for ABC and IHDP. For ETP, the effects are weaker—less than 1/2 of a standard deviation. Nevertheless, these effects are substantial compared to the short-term effects reported for Head Start and for the universal programs discussed in Sections 3.4 and 3.5, respectively.

In contrast to the IQ measures, the achievement measures used weight both cognitive and non-cognitive skill components more equally.[39] Achievement outcomes for ABC and PPP are strong. There is evidence of program effects on non-cognitive skills, but the different programs do not report strictly comparable measures. Furthermore, defining and measuring non-cognitive skills accurately is an open challenge that presents difficulties in detecting effects even when they are present.

Fadeout of Effects for Cognitive Skills

A general pattern for IQ and achievement test scores is that they tend to surge while children are in pre-K and then fade. In some cases, they completely dissipate. In two documented cases, IQ effects persist long after school entry: for the whole ABC sample (see Appendix D) and for some subgroups of IHDP (Duncan and Sojourner, 2013b). Even in those cases, the impacts during the program were stronger than the long-term impacts. All other studies in this paper that report the dynamics of impacts on test scores find that IQ or achievement gains dissipate. This is true for other demonstration programs (Weikart, 1970; Gray et al., 1982b), Head Start (see Deming, 2009; Zhai et al., 2014), and state programs (see Lipsey et al., 2013).

Figure 3.2a illustrates the fadeout phenomenon using evidence from PPP. IQ tests are usually scaled to show the level of a child relative to that of the overall population of their age. The decrease in standardized IQ for children in the treatment group after entering

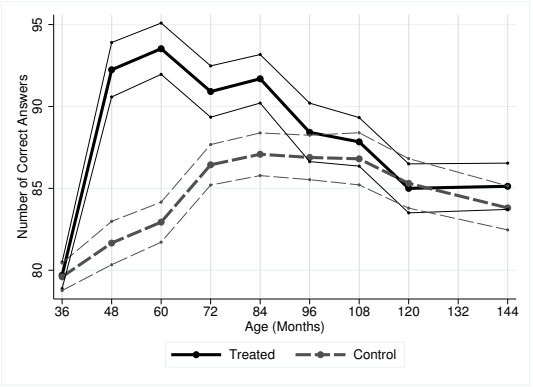
[39]See Heckman and Kautz (2012b).

elementary school indicates that the gap between them and an average US child increases. The figure does not reveal whether skills gained by the treatment group depreciate or those gained by the control group catch up. Figure 3.2b presents the raw scores in terms of total questions answered. They increase uniformly during childhood (Hojman, 2015). Additional figures illustrating the evolution of IQ and achievement scores over the life-cycle are presented for all programs in Appendix D.

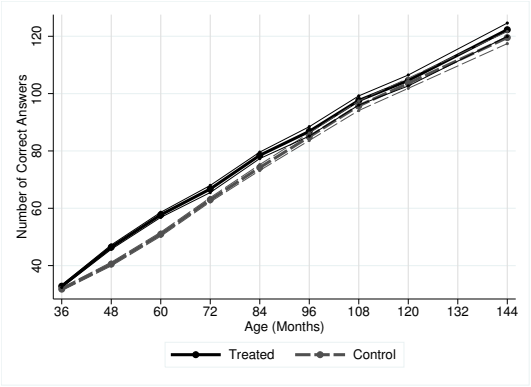
Hojman (2015) analyzes the causes of fadeout in cognition measured by IQ for PPP and ETP. He finds that the gains experienced by the treatment group occur rapidly during the first months of treatment and are followed by small or zero gains in the subsequent years of treatment. He also finds that almost all of the fadeout happens during the first year of elementary school. The gap between treatment and control groups narrows because the control group gains more from schooling. Measured IQ improves as a direct consequence of the initial formal educational experiences and the increase is roughly independent of the age at which entry into preschool or formal education begins. The laggard growth of IQ for all disadvantaged children may be consequences of the low quality of the schools they attend, the lack of stimulation in their home environments, or some combination of those factors. The precise causes are not known.

Figure 3.2: Dynamics of IQ in PPP

(a) Standardized IQ



(b) Raw IQ



Source: Reproduced from Hojman (2015). Note: The solid line represents the trajectory of the treated group, and the dotted line represents the trajectory of the control group. Thin lines surrounding trajectories are asymptotic standard errors. It shows standardized IQ as measured by the Stanford-Binet test in each year. IQ is age-standardized based on a national sample to have a US national mean of 100 points and standard deviation of 15 points. In Figure 3.2b, the scores are not standardized. The scores in it represent the raw scores, or the sum of the number of correct questions in each year.

Differences by Gender

A consistent finding across all four programs is the difference in treatment effects for males and females. This difference is substantial enough to create important gender differences in both benefit-cost ratios and internal rates of return for PPP and ABC. This pattern is consistent with the literature on differences in development between girls and boys.[40] Girls develop earlier. Uniform curricula across genders appears to benefit the laggard boys on many dimensions, but girls benefit as well, as we document in our discussion of the long-term treatment effects of ABC and PPP. In addition, all programs (except IHDP) target ages 3–4 when aggressive behavior that predicts adult aggression and participation in crime begins to manifest itself (White et al., 1994). Gender-specific curricula in preschool may be an appropriate strategy.

[40]Lavigueur et al. (1995); Kerr et al. (1997); Mâsse and Tremblay (1997); Nagin and Tremblay (2001); Bertrand and Pan (2011).

Treatment Effect Heterogeneity by Socio-Economic Status

IHDP served a more heterogeneous population compared to the other demonstration programs. A consistent policy-relevant finding for this program is the heterogeneity in treatment effects across socio-economic groups. The literature finds much higher treatment effects for the low-low birth-weight children ($\leq$ 2000 grams) when compared to the effects for the high-low birth-weight children ($>$ 2000 grams, $\leq$ 2500 grams).[41] For example, the effects on IQ at age 18 are negative but not statistically significant for the latter and are significantly positive for the former. Treatment effects are also heterogeneous by socio-economic status.

Brooks-Gunn et al. (1992) discuss the effects of the programs on IQ at age 3 and find that children whose mothers had a college degree or higher experienced no treatment effects on IQ, while children with relatively uneducated mothers had sizable effects. A recent study shows that program effects on IQ exhibit a gradient corresponding to household income, suggesting that poorer children experience the greatest benefits. Duncan and Sojourner (2013b) find that at age 2, the treatment effect for cognition accounts for .82 standard deviations for children of families with relatively low income with a standard error of .30, while the estimated effect is .46 for children of families with relatively high income with a standard error of .23.

3.3.5 Long-Term Outcomes

PPP and ABC are the only demonstration programs with follow-up during adulthood. A summary of their most important effects is given in Table 3.7, which is based on results from Heckman et al. (2010c, 2013b); Campbell et al. (2014), and García et al. (2015). The results reported in the table are statistically significant after accounting for multiple hypotheses testing across relevant, related outcomes. PPP caused a 56% increase in the high school graduation for females and a 29% increase in employment at age 40 for males. Other bene-

[41] Brooks-Gunn et al. (1994); McCormick et al. (2006).

ficial effects include criminal activity, employment, health behavior, and welfare take-up. In general, the table shows that PPP and ABC had statistically significant positive outcomes that persist into adulthood. Non-cognitive outcomes are notably absent due to lack of data. In PPP and ABC, and for early education programs in general, non-cognitive skills are not typically followed in the long term.

Table 3.7: Life-Cycle Outcomes, PPP and ABC

	PPP			ABC		
	Age	Female	Male	Age	Female	Male
Cognition and Education						
Adult IQ	-	-	-	21[c]	10.275	2.588
	-	-	-		**(0.005)**	(0.130)
High School Graduation	19[a]	0.56	0.02	21[c]	0.238	0.176
		(0.000)	(0.416)		**(0.090)**	**(0.100)**
Economic						
Employed	40[a]	-0.01	.29	30[c]	0.147	0.302
		(0.615)	**(0.011)**		(0.135)	**(0.005)**
Yearly Labor Income, 2014 USD	40[a]	$6,166	$8,213	30[c]	$3,578	$17,214
		(0.224)	(0.150)		**(0.000)**	(0.110)
HI by Employer	40[a]	0.129	0.206	31[b]	0.043	0.296
		(0.055)	(0.103)		(0.512)	**(0.035)**
Ever on Welfare	18–27[a]	-0.27	0.03	30[c]	0.006	-0.062
		(0.049)	(0.590)		(0.517)	**(0.000)**
Crime						
No. of Arrests[d]	$\leq$40[a]	-2.77	-4.88	$\leq$34[c]	-5.061	-6.834
		(0.041)	**(0.036)**		**(0.051)**	(0.187)
No. of Non-Juv. Arrests	$\leq$40[a]	-2.45	-4.85	$\leq$34[c]	-4.531	-6.031
One-sided permutation		**(0.051)**	**(0.025)**		**(0.061)**	(0.181)
Lifestyle						
Self-reported Drug User	-	-	-	30[c]	0.031	-0.438
-	-	-	-		(0.590)	**(0.030)**
Not a Daily Smoker	27[a]	0.111	0.119	-	-	-
		(0.110)	**(0.089)**	-	-	-
Not a Daily Smoker	40[a]	0.067	0.194	-	-	-
		(0.206)	**(0.010)**	-	-	-
Physical Activity	40[a]	0.330	0.090	21[b]	0.249	0.084
		(0.002)	(0.545)		**(0.004)**	(0.866)
Health						
Obesity (BMI >30)	-	-	-	30–34[c]	0.221	-0.292
	-	-	-		(0.920)	**(0.060)**
Hypertension I	-	-	-	30–34[c]	0.096	0.339
	-	-	-		(0.380)	**(0.010)**

Source: [a] Heckman et al. (2010c). [b] Campbell et al. (2014). [c] García et al. (2015). Note: This table displays statistics for the treatment effects of PPP and ABC on important life-cycle outcome variables. Hypertension I is the first stage of high blood pressure—systolic blood pressure between 140 and 159 and diastolic pressure between 90 and 99. "HI by employer" refers to health insurance provided by the employer and is conditional on being employed. [d] "No. of Arrests" includes offenses in the case of ABC, even where more than one offense was charged per arrest. For the further definitions of the outcomes, see the respective web appendices of the cited papers. Outcomes from Heckman et al. (2010c) are reported with one-sided $p - value$ which is based on Freedman-Lane procedure, using the linear covariates of maternal employment, paternal presence and SB (Stanford-Binet) IQ, and restricting permutation orbits within strata formed by a Socio-economic Status index being above or below the sample median and permuting siblings as a block. $p - value$s for the outcomes from Campbell et al. (2014) are one-sided single hypothesis constrained permutation $p - value$'s, based on the IPW (Inverse Probability Weighting) t-statistic associated with the difference in means between treatment groups; probabilities of IPW are estimated using the variables gender, presence of father in home at entry, cultural deprivation scale, child IQ at entry (SB), number of siblings and maternal employment status. $p - values$ for the outcomes from García et al. (2015) are bootstrapped with 1000 resamples, corrected for attrition with Inverse Probability Weights, with treatment effects conditioned on treatment status, cohort, number of siblings, mothers IQ, and the ABC high risk index.

3.3.6 Connecting Short-Term and Long-Term Effects

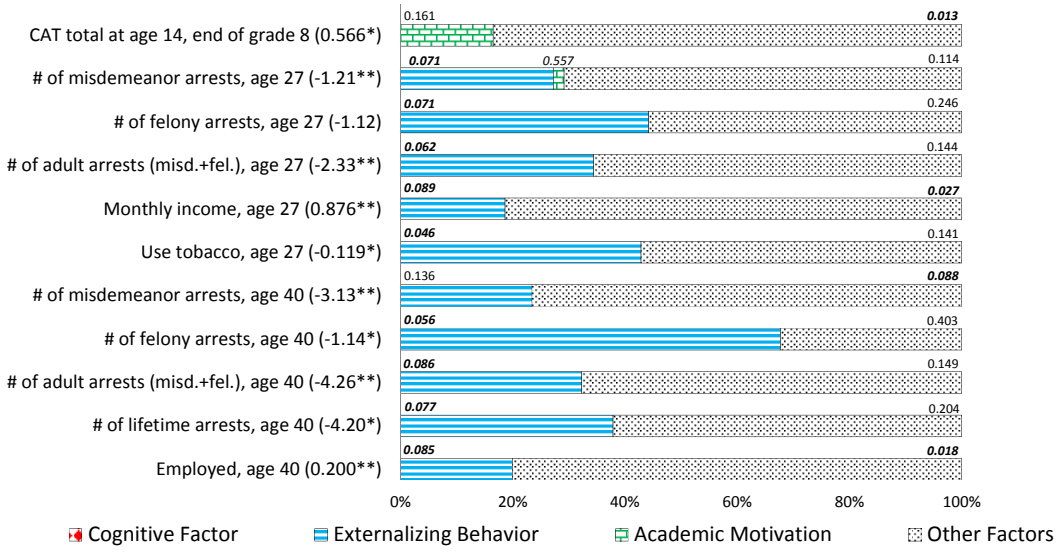
Dissipation of initial IQ gains is a common finding across programs. In some cases, IQ gains completely dissipate by the teenage years. Analysts focusing solely on IQ as a measure of program effectiveness confront a puzzle: Why do early childhood education programs have long-term effects if the effects on IQ dissipate? Heckman et al. (2013b) present a solution to this puzzle by considering the process through which skills form and develop. They find that program effects on non-cognitive skills are important determinants of later-life outcomes.[42] This conclusion highlights the importance of skill formation as a multi-skill dynamic process in which different skills complement each other.

Heckman et al. (2013b) decompose the effects of PPP on later-life outcomes using a mediation analysis. The results of this are reported in Figures 3.3 and 3.4.[43] They find that boosts in non-cognitive skills are substantial determinants of long-term effects. For females, academic motivation mediates 30% and 40% of the effects on achievement and employment, respectively. Further, reductions in externalizing behavior explain 65% of the reduction in lifetime violent crimes and reduce lifetime arrests and unemployment by 40% and 20%, respectively. There are persistent effects of boosts in non-cognitive skills even though in the short run, cognitive effects fade out.

[42] We use the term mediation analysis to refer to the exercise of decomposing effects of policies or programs on an outcome into distinct components. The outcome is usually thought of as an output and the components are the inputs generating this output. For a formal definition and analysis, see Heckman and Pinto (2015).

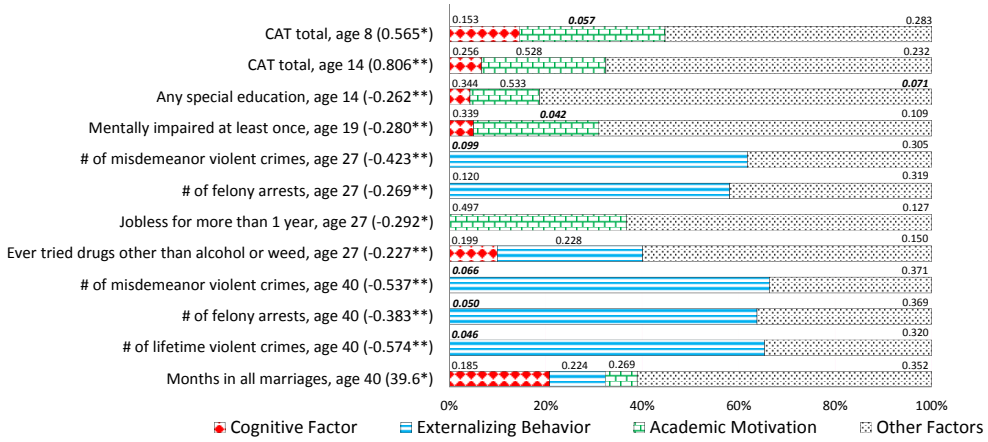
[43] See Heckman et al. (2013b).

Figure 3.3: Decompositions of Treatment Effects of PPP on Male Adult Outcomes



Source: Reproduced from Heckman et al. (2013b). Note: The total treatment effects are shown in parentheses. Each bar represents the total treatment effect normalized to 100 percent. One-sided $p-values$ are shown above each component of the decomposition. See the Web Appendix of Heckman et al. (2013b) for detailed information about the simplifications made to produce the figure. "CAT total" denotes California Achievement Test total score normalized to control mean 0 and variance of 1. Asterisks denote statistical significance: * – 10% level; ** – 5% level; *** – 1% level. Monthly income is adjusted to thousands of 2006 dollars using annual national CPI.

Figure 3.4: Decompositions of Treatment Effects of PPP on Female Adult Outcomes

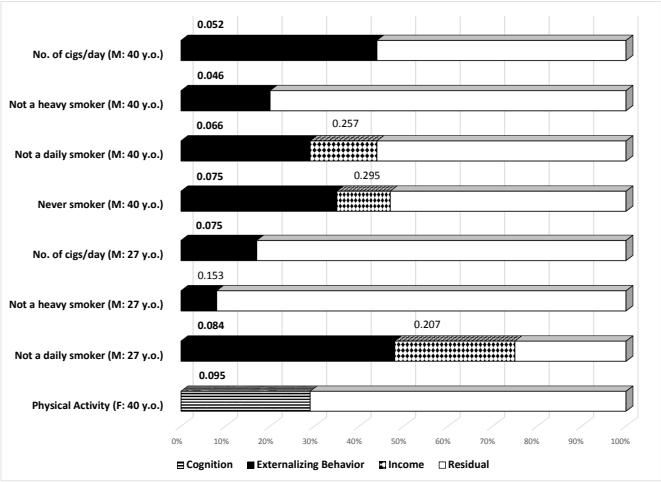


Source: Reproduced from Heckman et al. (2013b). See note in Figure 3.3.

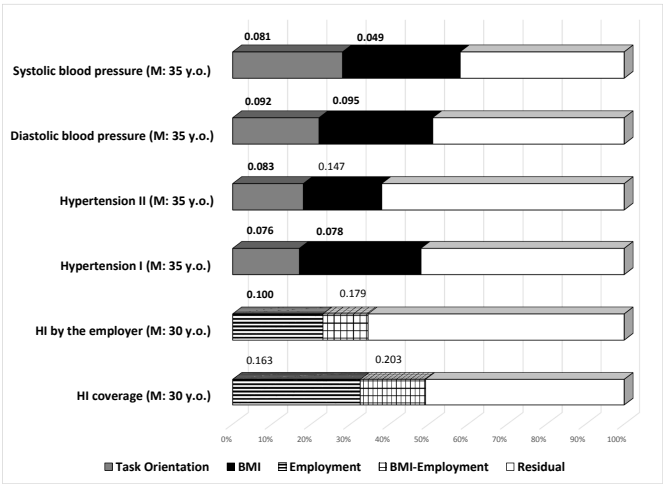
Conti et al. (2015) conduct a similar analysis for both PPP and ABC but focus on health outcomes. According to their findings, externalizing behavior is the primary mediator for the outcomes found in PPP, which is consistent with the findings in Heckman et al. (2013b). For ABC, they find that task orientation and childhood BMI mediate approximately half of the improvements in blood pressure and hypertension found for males in the treatment group. Figures 3.5a and 3.5b illustrate the results from their mediation exercises.

Figure 3.5: Decompositions of Treatment Effects of PPP and ABC on Male Adult Outcomes

(a) PPP



(b) ABC

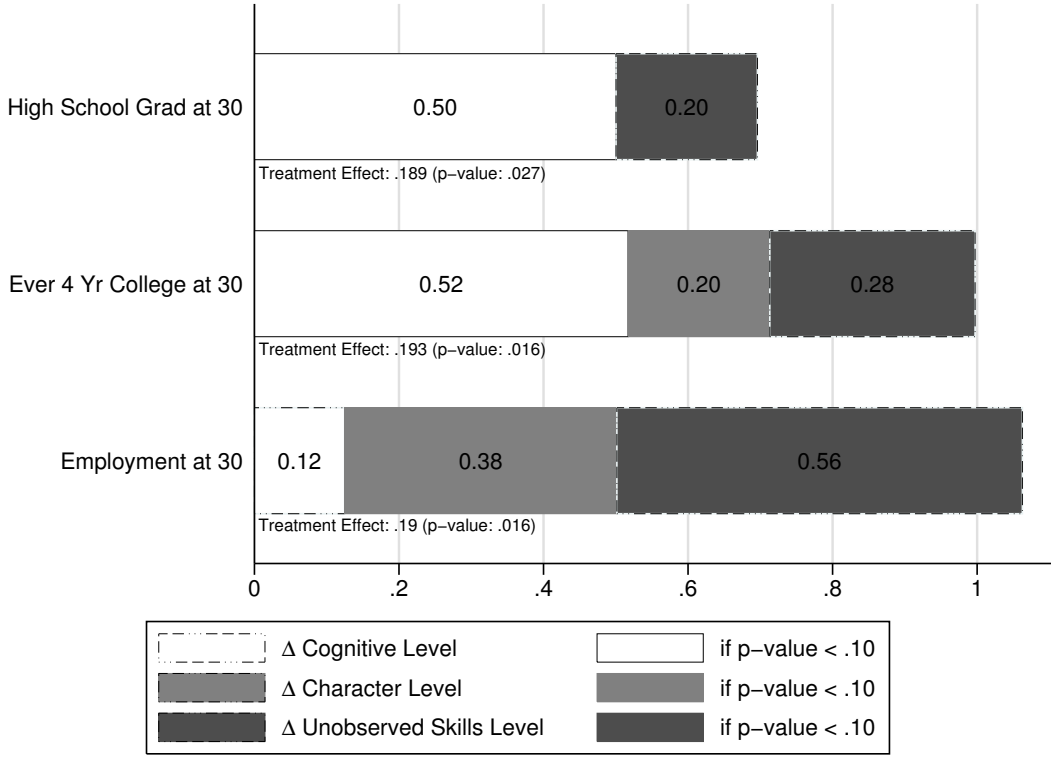


Source: Reproduced from Conti et al. (2015). Note: This graph provides a simplified representation of the results of the dynamic mediation analysis of the statistically significant outcomes for PPP and ABC. Each bar represents the total treatment effect normalized to 100%. One-sided $p - value$s that test if the share is statistically significantly different from 0 are shown above each component of the decomposition. The mediators displayed are: externalizing behavior, as in Heckman et al. (2013b) among the early childhood inputs; and income as in Heckman et al. (2010c) among the adult inputs. The complete mediation results and the definition of each outcome is reported in the Web Appendix of Conti et al. (2015). The sample the outcomes refer to (M = males; F = females) and the age at which they have been measured (y.o. = years old) are shown in parentheses to the left of each bar, after the description of the variable of interest. ***: significant at the 1% level; **: significant at the 5% level; *: significant at the 10% level.

García (2014) decomposes the ABC treatment effects pooling males and females. He analyzes three outcomes at age 30: high school graduation, ever being enrolled in a four-year college, and employment. See Figure 3.6. He shows that the more relevant the outcome is

for economic success, the less it is mediated through cognition and the more it is mediated through non-cognitive skills.

Figure 3.6: Decompositions of Treatment Effects of ABC on Male and Female (Pooled) Adult Outcomes



Source: Own calculation. Note: This plot decomposes the total treatment effect ABC has on graduating high school, ever enrolling in a four year college, and employment at age 30. The figure presents the components of Laspeyres decomposition of the relevant outcome on a measure of cognition and a factor summarizing character skills. Cognition is measured at age 21 using the Woodcock-Johnson Test of Achievement. Character is measured at age 15 by a factor created using measures of conscientiousness. The numbers inside the bars represent the proportion explained by each component. They do not sum to 1, because the decompositions condition on socio-demographic variables which are not displayed above. See García (2014) for more details.

3.3.7 Cost-Benefit and Rate of Return Analyses

Cost-benefit and rate of return analyses produce concise, policy-relevant statistics for assessing the social benefits of programs. While there is a vast literature evaluating treatment effects for demonstration programs, cost-benefit analyses are scarce (Currie, 2001b). This

scarcity arises from the difficulty in securing the relevant data. Cost-benefit analyses require comprehensive data in order to account for impacts over the life-cycle. Very few programs have been evaluated rigorously using cost-benefit analysis. In fact, only PPP and ABC have the data required to conduct such exercises, accounting for the variety of outcomes including criminal activity, income, and health.

Heckman et al. (2010d) substantially improve on an earlier cost-benefit analysis of PPP by Belfield et al. (2006) that does not report standard errors, does not disaggregate by gender, and uses an *ad hoc* method for forecasting out of sample earnings gains. Heckman et al. (2010d) use a broader base of data and substantially refine the estimates in Belfield et al. (2006). Both papers incorporate costs of education and estimates of benefits. Heckman et al. (2010d) additionally account for the deadweight loss created by collecting public funds. They calculate standard errors for their estimates. They invoke standard assumptions about the deadweight losses associated with collecting tax revenue to support programs, the social costs of crime, and the procedures used to extrapolate future benefits. The range of estimates for the annual rate of return pooled across genders is 7-10% per annum. The corresponding range for the benefit-cost ratio is 3.9-6.8. Disaggregating by gender produces higher estimates. All of these estimates are statistically significant. Their preferred estimates are presented in the columns under "PPP" in Table 3.8.

García et al. (2015) present the benefit-cost analysis of ABC through age 35.[44] Their study demonstrates the social efficiency of this program. The benefit-cost estimates are lower when compared to PPP, in part because the costs of the program are higher. It is the first study to account for life-cycle gains in health using age 34 biomarkers to project future health. Other important sources of benefit from the program are gains in parental income while participants are young, gains in later-life income, and decreases in criminal activity. The study finds an overall benefit-cost ratio of 3.2:1 and an internal rate of return of 11%.[45] When decomposed by gender, the results are much stronger for males because

[44]This paper extends the methodology in Heckman et al. (2010d).

[45]The estimates are statistically significant at the 10% level.

the main benefits are reduced criminal activity and improved health, both of which show stronger effects for males.[46]

Table 3.8 displays the main components of the cost-benefit analyses of PPP and ABC. Lifetime earnings and health benefits are crucial components of the benefits of ABC, as well as reductions in criminal activity corresponding to serious crimes for males (García et al., 2015).[47]

Gains in parental income are an important component of the returns to ABC because the program provided care for up to nine hours a day, thus enabling mothers to increase their labor supply. Early childhood education has effects not only on the children, but also on the economic lives of their families. It is a form of enriched childcare that enables mothers to work and to provide additional resources for disadvantaged families. There are likely intergenerational effects on the children of participants in both programs as well. Data being collected on PPP will enable analysts to compute the gains to the children of participants (Heckman, 2015).

Our evidence on the social benefits of ABC and PPP does not suggest that these programs should be slavishly imitated. It suggests guiding principles for future policy which can only benefit from the knowledge acquired since the time these programs were implemented. It shows the promise of such programs and provides a lower bound on what is possible.

[46]Barnett and Masse (2007) provide an estimate of the benefit-cost ratio for ABC of 2.5:1, but give no standard error for their estimate, do not aggregate by gender, and use an *ad hoc* method to forecast future benefits of treatment. Their calculation does not account for the most recent follow-up of ABC, including the substantial boost in health of participant males. Its main components are gains on parental income when the children are young and individual income up to age 21, but their estimates of earnings impacts are not credible.

[47]Health data were not collected for PPP.

Table 3.8: Costs and Benefits of PPP and ABC, 2014 USD

	PPP			ABC		
Net Present Value	Female	Male	Pooled	Female	Male	Pooled
Parent Income[a]	-	-	-	$88,358	$88,358	$88,358
Control Group Preschool[b]	-	-	-	$1,832	$1,292	$1,469
Program Cost per Recipient[c]	$31,168	$31,168	$31,168	$91,519	$91,519	$91,519
Education Costs[d]	$9,626	$(19,678)	$(7,528)	$28,715	$5,083	$12,586
Subject Labor Income[e]	$149,157	$50,269	$91,272	$36,270	$89,417	$70,798
Subject Transfer Income[f]	$9,656	$4,248	$6,490	$2,614	$1,729	$2,256
Savings in Medical Expenditures[g]	-	-	-	$9,920	$22,236	$19,604
Savings in Crime[h]	$26,400	$131,330	$87,823	$9,924	$219,911	$101,726
Quality of Life (QALY) Benefits[i]	-	-	-	$2,997	$21,845	$19,985
Net Benefit	$144,420	$174,358	$161,944	$31,671	$358,352	$200,009
Benefit-Cost Ratio	7.3:1	5.4:1	6.6:1	1.4:1	4.9:1	3.2:1
S.E.	(3.2)	(3.0)	(2.7)	(0.98)	(3.19)	(1.53)
Internal Rate of Return (%)	9.5	9.7	7.7	4.1	12.7	11
S.E.	(2.7)	(3.0)	(2.6)	(0.10)	(0.06)	(0.05)

Source: PPP estimates from Heckman et al. (2010d); ABC estimates from García et al. (2015). Note: PPP results use a 3% discount rate, and ABC results use a 4% discount rate. All results take into account deadweight loss of public spending of 50%. Cost-benefit ratios in PPP do not exactly reflect the net benefits and costs, because the ratios and the internal rates of return are adjusted for compromised randomization. [a] Parental income: annual labor income during children's ages 0 to 15. [b] Costs incurred by parents of the control group children for sending them to preschool. [c] Cost per recipient of either PPP or ABC. [d] Education costs from elementary school up to latest education over the life-cycle. [e] Labor income from ages 21 to 65. [f] Total income transferred from the government to the individual. Given this is a transfer from one agent of society (government) to another (individual), this number only accounts for the deadweight loss generated by the transfer. [g] Total medical expenditures from age 34 up to expected death. Treatment group individuals spend more, on average, because they live longer due to positive treatment effects on multiple health measures. [h] Savings due to crime reduction, accounting both for costs to victims and prison costs. [i] QALY stands for quality-adjusted life years. Quality of life is measured by an index of activities of daily life and takes values between 0 and 1, where 0 represents death and 1 represents full health. Each year of life is valued at $150,000 and weighted by the quality of life. Standard errors are obtained using bootstrapping.

3.3.8 Summary of the Evidence from Demonstration Programs

The evidence on demonstration programs supports several general conclusions. High-quality early childhood education programs targeted to disadvantaged children have long-term positive effects on important social and economic outcomes. Although the short-term effects on IQ tend to fade, a careful examination of program effects on multiple skills and dynamic skill formation demonstrates how improvements in non-cognitive skills generate lasting effects on many later-life outcomes. The strong estimated effects and the evidence on social efficiency supported by cost-benefit analyses provide a strong case for the public provision of high-quality targeted programs. These programs also provide childcare and facilitate working by the mothers of disadvantaged children.

3.4 Evidence from Head Start

Head Start is the largest and oldest public early childhood education program in the US.[48] Evidence on it is important for understanding the benefits of early education. There are multiple evaluations of Head Start based on different methodologies and data sources. Studies use evidence from both nationally representative datasets and a randomized controlled trial designed to evaluate Head Start.[49]

The evaluations of Head Start report contradictory evidence, in part because they fail to articulate the different policy questions that they implicitly answer. Ohio University and Westinghouse Learning Corporation (1969) and McKey et al. (1985) are two highly-cited studies claiming to find no long-term effects on relevant socio-economic outcomes. On the other hand, Ludwig and Miller (2007) and others claim that the program recovers its costs and then some through the gains it creates in the educational attainments of participants.

[48] Other large-scale, targeted early childhood education programs in the US include the Chicago Parent-Child Centers and Early Head Start. Reynolds and Temple (1998, 2006), Reynolds et al. (2011), and Love et al. (2005) respectively evaluate them. Reynolds refuses to release his full data set, so it is impossible to verify his claims.

[49] The Head Start Impact Study (HSIS) is reported in Puma et al. (2012).

As a group, these studies are imprecise about the counterfactuals being estimated. They typically do not discuss the alternative childcare arrangements available to participants at the time they were enrolled. This section presents evidence from evaluations with rigorous methodologies. We discuss studies that address well-defined policy questions that consider the availability of alternative childcare arrangements. These studies find that Head Start has positive effects in the short term on measures of cognitive and non-cognitive skills. They are reinforced by the evidence from several studies evaluating long-term outcomes, using many different datasets and methodologies, all of which find impacts in substantive adult outcomes.

3.4.1 Overview of Head Start

Head Start is a means-tested, federal preschool program founded in 1965. It is the largest ongoing early childhood education program in the US. Children aged 3 or 4 are eligible if family income is below or at the poverty line (though there is a designated quota for children whose families are above the poverty line). Children who enter the program at age 3 receive two years of treatment, which is mainly given in center-based programs. Its objective is to foster cognitive and non-cognitive development and school-readiness with a "whole child" approach. It pursues these objectives by granting funds to qualified centers. In turn, these centers are required to maintain high performance standards.

Performance standards within Head Start mandate minimal quality levels for health, nutrition, and family partnerships. Head Start centers must verify the child's health status and screen for behavioral or mental health problems. Head Start centers also provide services to parents and families in order to improve the "whole" environments of the children.[50]

Despite its uniform minimum standards, there is substantial heterogeneity in the quality of Head Start centers, both in services and in the skills of the staff. While many categorize Head Start as a high-quality program, we cannot make an absolute judgment of "the" effect

[50] Administration for Children and Families, Office of Head Start (2009).

of Head Start due to the substantial heterogeneity in treatment effects.

Early Head Start

Early Head Start is an offshoot of Head Start. Established in 1994, it serves pregnant women and children under age 3 who meet Head Start's income eligibility criteria. All Early Head Start programs offer full-day, full-year treatment and have center-based and/or home-visiting components. Like Head Start, it has a "whole child" approach with the goal of preparing children for future growth and development. Notably, it focuses on nurturing healthy attachments between children and their parents and caregivers. Both Early Head Start and Head Start offer transition services to help children adjust and move smoothly from Early Head Start to Head Start and from Head Start to kindergarten. We do not review results from Early Head Start due to the scarcity of rigorous evaluations of it, their short-term follow-up, and high heterogeneity of the treatments offered.[51]

Comparability with Demonstration and Universal Programs

Like the demonstration programs previously discussed, Head Start is means-tested and provides services beyond center-based care. In fact, Head Start shares important features with PPP and ABC, including curricular and extracurricular program components. There is a relationship between Head Start and previous early childhood education programs, such as PPP and ABC. Roughly 30% of the Head Start Impact Study (HSIS) centers use the *High-Scope* curriculum, which was developed from the PPP curriculum. This curriculum seeks to improve school-readiness by targeting age-appropriate developmental tasks such as gross/fine

[51] One evaluation of Early Head Start is by Love et al. (2005). They use an instrumental variable approach to assess the effects of program participation on a variety of outcomes at age 3. Early Head Start had three types of implementations: (i) center-based programs; (ii) home-based programs; and (iii) mixed approach programs. When pooling the sample, they find important gains on mental development, cognition, and some measures of child behavior. Unfortunately, the results are not as clear when the samples are broken down into type of implementation. The available Early Head Start evaluations do not isolate the effects by treatment stream. Furthermore, it fails to provide estimates of the effects of the program in the long-term because data are not available. Given its similarities with Head Start, future evaluations should discuss whether control contamination is an issue.

motor, language and literacy, cognitive, and social-emotional development. It emphasizes the importance of a supportive learning environment and the relationship between caretaker and child.[52] Second, ABC and Head Start share extracurricular components, including medical and nutritional services. 88% of the children who participated in HSIS received nutritional services through the program. Some 80% received medical services. ABC and Head Start also share operational similarities (Puma et al., 2012). 45% of Head Start centers offer care from birth to age 5 by combining Head Start and Early Head Start.[53] Further operational similarities include access to full-day care and transportation to the center. 68% of children who participated in HSIS were offered the option of attending full-day care, and 63% had the option of being transported to the center, as in ABC.[54]

Head Start also has similarities with the universal programs we discuss in Section 3.5. It is a wide-ranging program that serves diverse disadvantaged populations. Analyses of Head Start are not subject to questions of large-scale reproducibility that burden the evidence from demonstration programs.

3.4.2 Data

There are two sources of evidence on Head Start: (i) HSIS, which is the largest randomized control trial on early childhood education in the US; and (ii) studies based on nationally representative observational data, such as the Panel Study of Income Dynamics (PSID; see Panel Study of Income Dynamics, 2015), the National Longitudinal Survey of Youth 1979 (NLSY79; see Bureau of Labor Statistics, 2015), and the Children of the National Longitudinal Survey of Youth (CNLSY; see Bureau of Labor Statistics, 2011), which record participation in Head Start and have long-term follow-up data. As the largest randomized control trial of an early childhood education program in the US, HSIS is a preferred source of data for analysts. It does not suffer from the small sample size problems that plague demon-

[52] Puma et al. (2012).
[53] Administration for Children and Families, Office of Head Start (2014).
[54] Puma et al. (2012).

stration programs. Moreover, it is nationally representative of Head Start centers across the nation, which implies generalizability of its results. Yet, it suffers from some major limitations that complicate the estimation of meaningful policy parameters, namely: heterogeneous treatments across centers, lack of long-term follow-up, and control contamination.

Heterogeneous Populations and Treatment Alternatives

Head Start provides funding to local centers, which attempt to tailor treatment of the problems of the populations they serve. Thus, the quality of the centers, the populations served, and the alternatives available to parents vary among centers.

Lack of Long-Term Follow-Up

HSIS has follow-up until age 9 and cannot be used to evaluate long-term effects of Head Start. Lack of long-term follow-up in HSIS is mitigated by the availability of long-term outcomes in nationally representative data such as the PSID, NLSY79, and CNLSY. However this results in an additional limitation on evaluations of Head Start, as long-term evaluations need to address the methodological challenges of integrating non-experimental data with experimental data.

Control Contamination

An important challenge emerges from the extensive control contamination that is present in HSIS. While the control group was denied treatment in the study centers—that is, the centers participating in HSIS—nothing prevented control (or treatment) families from seeking alternative options. This alternative could even include other centers providing Head Start. In fact, 15% of the control group attended other Head Start centers. In the HSIS study, some 40% of the control group used center-based care. Therefore, estimates of treatment effects that do not account for control contamination compare Head Start to Head Start for many participants. Such estimates—unsurprisingly—are close to zero and do not speak to

the efficacy of Head Start compared to the home care provided by parents.

We present short-term and long-term evidence on the impacts of HS in the following section. We summarize the evidence from all sources in Table 3.9.

3.4.3 Short-Term Outcomes

Puma et al. (2012) report a battery of mean differences between the treatment and "control" groups followed in HSIS using data through the age 9 follow-up. They report estimates for an age 3 cohort and age 4 cohort. The age 3 cohort received at least one year of treatment; after the first year of treatment, 63% of the treatment group remained at a Head Start center, and 26% of the treatment group were in some other center-based care arrangement. The age 4 cohort received only one year of treatment. For both cohorts, they report short-term positive effects for most measures of cognition which disappear by age 9. There are some treatment effects for non-cognitive skills, but the measures used are unreliable.[55] There are positive effects on parenting quality, especially for the age 3 cohort. Parents of the age 3 cohort spanked their children 14% less than control parents after the first year of treatment; by the age 6 follow-up, they spanked their children 9% less. The authors report that these estimates are significant at the 10% level but do not report exact p-values or standard errors. The control group had access to early childhood education alternatives, including other Head Start centers, so the reported treatment effect does not compare Head Start to home-based childcare.

Ludwig and Phillips (2008) use cognitive outcomes measured at the end of the first year of treatment and attempt to improve the interpretation of the estimates by statistically adjusting for the presence of control children who attend a Head Start center not in the HSIS study. To account for differences in enrollment to Head Start in the treatment and control group, they use a Bloom (1984) estimator to adjust intent-to-treat estimates reported in

[55]Treatment effects on the same measures of non-cognitive skills vary in sign depending on whether the measure was parent- or teacher-reported. Parent-reported measures yield favorable treatment effects, while teacher-reported measures yield unfavorable treatment effects.

Puma et al. (2005). They find effect sizes of .346 for the age 3 cohort with standard error .074 and an effect size of .319 for the age 4 cohort with standard error .147.[56] Their study does not address control contamination of other types. These estimates can be understood as estimates of the effects of offering Head Start in one center: the impact of Head Start at the center against the next best alternative which may be another Head Start center. When considering the effectiveness of providing public early childhood education programs compared to no programs at all, it is not the policy-relevant parameter.

Two recent studies address control contamination in HSIS more systematically. They relate their estimates to theoretical parameters in order to answer well-defined and relevant policy questions.[57] Both studies provide estimates of the average treatment effects in Head Start compared to different alternatives available to parents: (i) other preschool programs; and (ii) home care. Their estimates are based on five exhaustive and mutually exclusive groups: (i) those who are always Head Start users (11%); (ii) those who are always preschool users (11%); (iii) those who always keep children at home (12%); (iv) those who enroll in Head Start[58] (20%); and (v) those who stay at home after randomization into the program (45%).[59]

Identification in both papers relies on strong functional form assumptions. Feller et al. (2016) use a version of the standard econometric selection model and rely heavily on normality assumptions on the observed variables driving selection into treatment to identify their reported treatment effects. Kline and Walters (2014) present a much richer interpretive framework but rely on normality to characterize dependence among choices and outcomes, although they do not impose normality on the full model as do Feller et al. (2016). These studies discuss the identification problems present when using a single randomization to identify the effects of multiple choices.[60]

[56] Literacy is measured by the Woodcock-Johnson letter identification test.

[57] Feller et al. (2016); Kline and Walters (2014).

[58] "Compliers" in the language of LATE.

[59] We take these numbers from Feller et al. (2016). Kline and Walters (2014) report very similar percentages.

[60] See Heckman and Vytlacil (2007) for a general analysis of multiple competing choices and the use of

Both papers give estimates of the effect of Head Start relative to staying at home, which is the closest estimate of the parameter assessing the effect of Head Start relative to no treatment at all. The magnitudes of their preferred estimates on cognition are different: 0.23 of a standard deviation in Feller et al. (2016) (standard error .038) and 0.38 of a standard deviation in Kline and Walters (2014) (standard error .047).[61] Kline and Walters (2014) find negative selection into the program. Individuals who gain the most are the least likely to participate. After correcting for selection, the average treatment effect on the population is as high as 0.47 standard deviations of test scores (standard error .110), which approaches the effect that demonstration programs have on early measures of cognition. Both papers conclude that the effect of Head Start is similar to that of the alternative, local, center-based preschool alternatives and are both better than home care. This underscores the importance of carefully defining the alternative against which Head Start is compared.

Another recent study (Zhai et al., 2014) uses HSIS data to evaluate the short-term effects of Head Start. They compare individuals assigned to the treatment group with individuals assigned to the control group. The control group received care from three alternatives: (i) parental care; (ii) care from relatives; and (iii) care from another center. For comparison, they match individuals in the treatment group to three subsamples of the control group using standard methods for controlling for selection on observables.[62] They assess measures of both cognitive and non-cognitive behavior, as reported by the parents. Their findings on cognition are similar to the findings of Feller et al. (2016) and Kline and Walters (2014). They find that children who would have been cared for by their parents or relatives benefit the most from Head Start. The effects sizes on PPVT are .30 (parental care) and .19 (care from relatives) points at age 3 and .15 (parental care) and .30 (care from relatives) points at age 4, for the respective comparison groups. The evidence is somewhat ambiguous on program

instruments in this context.

[61] One of the reasons for this discrepancy is the use of different measures of cognition. Feller et al. (2016) use the Peabody Picture Vocabulary Test (PPVT), while Kline and Walters (2014) use an index of various measures.

[62] Inverse probability weighting.

effects for non-cognitive outcomes, but using parent reports, children generally become less aggressive and hyperactive at ages 3 and 4.[63] Teacher-reported measures of non-cognitive outcomes have negative treatment effects (see Puma et al., 2012). Zhai et al. (2014) do not report standard errors for their estimates.

3.4.4 Long-Term Outcomes

HSIS has no long-term follow-up. Evaluating the long-run impacts of Head Start requires use of non-experimental methods. We present results from such methodologies and discuss their policy implications.

Currie and Thomas (1995b), Garces et al. (2002), and Deming (2009) use longitudinal data in conventional, but controversial, panel data "fixed-effects" models that assume that the unobserved characteristics driving selection into treatment—and into preschool in general—are constant across time and are identical across children within families. They control for access to alternative early education programs to address the problem of control contamination.

Currie and Thomas (1995b) find short-term effects on cognition for both African-American and white children. However, these gains fade out for African-American children. Deming (2009) finds short-term effects for African-American but not for white children, and also finds a fadeout pattern consistent with that reported in Currie and Thomas (1995b). These studies are inconclusive about the effectiveness of the program because they do not consider their benefits on the multiple skills known to be important predictors of life outcomes.

Garces et al. (2002) and Deming (2009) measure treatment effects on outcomes during adulthood. Both studies find positive effects on high school completion and college attendance—the former for white enrollees and the latter for African-American enrollees.

[63] Bitler et al. (2014) present evidence relevant to our discussion using quantile instrumental variable methods. Children with relatively low skill endowments or from disadvantaged backgrounds benefit the most from treatment in Head Start. A serious limitation of these methods is the assumption of rank preservation in treatment and control distributions. When tested, this assumption is usually rejected. (See, e.g., Cunha et al., 2005 and Kline and Tartari, 2015.)

Garces et al. (2002) document positive effects on crime for African-American participants, but Deming (2009) finds no effects on crime. Although these studies attempt to account for selection into treatment, they only allow for a single additive unobserved component generating selection within the family and across time. Therefore, they cannot determine if the differences in their results are due to heterogeneity in treatment, problems in the specification of the models, differences in the populations, or something else.

Ludwig and Miller (2007) exploit variation in access to technical assistance for implementing Head Start in 300 poor counties, offered by the Office of Economic Opportunity in the 1960s. These counties were 50–100% more likely to participate in Head Start when compared to similarly situated counties. They find no notable differences in baseline characteristics between their 300 poor counties and their comparison counties. The authors find that Head Start has beneficial effects on mortality and schooling, although these findings are, at best, suggestive because they are based on limited data. Their reported effects are identified by comparing the outcomes in the 300 poor counties with other poor counties where alternatives to early childhood education are very limited. Their evidence is consistent with the finding that treatment is especially effective for disadvantaged children.

In the best available study, Carneiro and Ginja (2014) examine the long-term effects of Head Start by exploiting discontinuities in eligibility rules using the NLSY79 (Bureau of Labor Statistics, 2015) and the CNLSY79 panel data sets. They show that there are multiple eligibility thresholds across years, states, family size, and family structure. This distinguishes their study from standard regression discontinuity designs with a single threshold. They estimate the marginal effect of relaxing eligibility requirements for different groups of the population. This methodology is important when relating their findings to policy questions because it allows for comparison of the effects across individuals with different alternatives.

The authors report long-term positive effects on health behaviors, such as the number of visits to the doctor, use of medicine, and reduced smoking, as well as on behavioral outcomes, such as grade repetition and special education. They also find that the program

reduces obesity at ages 12 and 13, depression and obesity at ages 16 and 17, and crime at ages 20 and 21. As in the case of demonstration programs, Head Start is judged to be effective when it is evaluated using multiple outcomes, rather than focusing solely on cognitive outcomes.

3.4.5 Cost-Benefit Analyses

Although a formal cost-benefit analysis for Head Start is not available, several studies present limited calculations of the social benefits of the program. Currie and Thomas (1995b) find that effects on African-American enrollees are not sufficient to recover the costs of the program, while the results for whites are sufficient to do so. Ludwig and Miller (2007), Deming (2009), Kline and Walters (2014), and Carneiro and Ginja (2014) argue that the social returns of the program are positive. They do not account for many relevant benefit components and interpret their results as lower bounds. We consider this evidence as, at best, suggestive, since it is based on rough calculations and approximations and therefore is less definitive than the evidence on effectiveness from the demonstration programs. Nonetheless, it is consistent with their estimate. An example of this sort of analysis is the study by Kline and Walters (2014), who use the estimated effects reported for the Tennessee Star Study on earnings to link the short-term effects on cognition to earnings in Head Start.[64] Their calculation is, at best, approximate, because the programs have different objectives and did not serve comparable populations.[65]

[64]The earnings estimates for their calculations come from Chetty et al. (2011).

[65]This practice is widely used in the literature. Many of the current analyses of the long-term gains generated by early education use *ad hoc* relationships between short-term measurements and long-term outcomes to forecast future gains from the program (see Barnett and Masse, 2007 and Bartik et al., 2012), a practice of questionable value. García et al. (2015) present a more principled extrapolation analysis and a discussion of general procedures.

Table 3.9: Evidence Across Studies of the Impacts of Head Start

Study	Currie and Thomas (1995b)	Garces et al. (2002)	Ludwig and Miller (2007)	Deming (2009)	Carneiro and Ginja (2014)	Feller et al. (2016)	Kline and Walters (2014)	Zhai et al. (2014)	Abecedarian (Various sources)	Abecedarian (Various sources)
Dataset	C-NLSY	PSID	Multiple	C-NLSY	C-NLSY	HSIS	HSIS	HSIS		
Subpopulation	AA	AA, mother edu. $\leq$ edu. $\leq$ high school		AA	Males				AA, low child IQ at entry & SES	98% AA, low mother IQ, & low SES
Years of birth	1979-1987	1966-1977	1960-1975	1979-1986	1977-1996	1998-1999	1998-1999	1998-1999	1959-1964	1972-1977
Impacts										
IQ/achievement, ages 3-4	-	-	-	-	-	**0.230**	**0.375**	**0.30**[a]	-	**0.880**[b]
	-	-	-	-	-	(0.038)	(0.047)	-	-	(0.147)
Behavior, ages 3-4	-	-	-	-	-	-	-	**0.35-0.19**[a]	-	-
	-	-	-	-	-	-	-	-	-	-
IQ/achievement, ages 5-6	**0.46**	-	-	**0.287**	-	-	-	-	**0.763**[c]	**0.427**[c]
	(0.129)	-	-	(0.095)	-	-	-	-	(0.127)	(0.227)
IQ/achievement, ages 7-21	0.201	-	-	0.031	-	-	-	-	**0.084**[c]	**0.300** [c]
	(NA)	-	-	(0.076)	-	-	-	-	(0.059)	(0.213)
Grade retention ever	-0.008	-	-	**-0.107**	-	-	-	-	-	**-0.244**[b]
	(0.098)	-	-	(0.056)	-	-	-	-	(0.151)	
High School Grad. (no GED)	-	0.00	0.117	0.067	-	-	-	-	**0.56** [d]	**0.185**[b]
	-	(0.071)	(0.080)	(0.044)	-	-	-	-	(0.093)	(0.210)
Attended some college	-	0.031	**0.028**	**0.136**	-	-	-	-	-	-
	-	(0.067)	(0.019)	0.049	-	-	-	-		-
Earnings, ages 23-40	-	0.051	-	-	-	-	-	-	$6,166[d]	$8,499[b]
	-	(0.357)	-	-	-	-	-	-	(8244)	(8018)
Idle	-	-	-	-0.030	-	-	-	-	-	-
	-	-	-	(0.053)	-	-	-	-	-	-
Ever booked crime	-	**-0.126**	-	0.051	-	-	-	-	**-2.77**[d]	**-5.739**[b]
	-	(0.05)	-	0.050	-	-	-	-	(1.590)	(4.250)
Behavior Index, ages 12-13	-	-	-	-	-0.647	-	-	-	-	-
	-	-	-	-	(0.582)	-	-	-	-	-
Depression Scale, ages 16-17	-	-	-	-	-0.552	-	-	-	-	-
	-	-	-	-	(0.489)	-	-	-	-	

Note: Impacts are in bold whenever they would be significant in a t-test at the 10% significance level. SES stands for sociso-economic status. Impacts on IQ/achievement scores are reported in standard deviations. Currie and Thomas (1995b) originally report impacts on IQ/achievement in terms of test scores: PPVT at age 8 in Currie and Thomas (1995b) is calculated using their interaction of Head Start and Peabody Picture Vocabulary Tests coefficient. The SE for the predicted impact at this age is not reported. Our calculations use bootstrapped standard errors. Grade retention is measured at age 5 in Currie and Thomas (1995b) and at age 18 in all other studies. Earnings in Garces et al. (2002) are measured in logs. Ludwig and Miller (2007) use census data, Vital Statistics, and the NELS. For the sake of brevity, we limit the number of estimates we present from Ludwig and Miller (2007) to only one per data set: the impact of treatment on mortality is from the Vital Statistics, impact on high school completion is from the NELS, and impact on attending some college is from the census. Impact on high school completion and college attendance are for children roughly 18-24 years old. Feller et al. (2016) originally reported 95% posterior intervals of 0.15, 0.30 during the Head Start Program. Impacts reported in Kline and Walters (2014) are estimated from a summary index created from Peabody Picture Vocabulary Tests and Woodcock-Johnson III Preacademic Skills tests taken in Spring 2003; this index is standardized to have mean 0 and a standard deviation of 1. The Center for Epidemiological Studies Depression Scale in Carneiro and Ginja (2014) measures symptoms of depression in percentile scores, where higher scores are negative. AA: African-American. [a] For IQ in Zhai et al. (2014), we report effect sizes on PPVT at ages 3 and 4 (they coincide). For behavior we report hyperactiveness at these same ages. Only Zhai et al. (2014) accounts for multiple hypotheses testing, across similar outcomes. For the studies using HSIS data, all treatment effects are reported in terms of effect sizes and, thus, are comparable across studies. For the estimation results that are reported separately for 3-year-old and 4-year-old cohorts, we use simple averages. For ages 3–4, we report the results in Feller et al. (2016), Kline and Walters (2014) and Zhai et al. (2014), measured after the Head Start year. For ages 5–6, we report the results in Zhai et al. (2014) measured after the children finish kindergarten. The comparable results in Puma et al. (2012) are 0.135 for ages 3–4 and 0.085 for ages 5–6. [b] Impacts are reproduced from the Web Appendix for García et al. (2015). IQ is reported at age 3 using the Stanford-Binet Intelligence Scale. Grade retention is reported for K-12 schooling. High school graduation is reported at age 19. Income is reported at age 30 in 2014 dollars. "Ever booked crime" represents total arrests by age 34. [c] Own calculations. See Table 3.4; impacts are in bold whenever they have a significant one-sided, permutation $p - value$. IQ for ABC is reported at age 5 and 8 using the Wechsler Intelligence Scale. [d] Results taken from Table 3.7; see the corresponding table note for details. This table only displays results for females from PPP. "Ever booked crime" represents total arrests by age 40.

3.4.6 Summary of the Evidence from Head Start

We summarize the estimates for Head Start that are reported in the literature in Table 3.9. As previously noted, the counterfactuals identified in these studies are not clearly specified. We also present comparable estimated effects from PPP and ABC by way of comparison. The effects reported in demonstration programs are typically stronger.

It is important to note that: (i) the studies based on HSIS only evaluate the impact of a single year of Head Start; (ii) the Head Start population is less disadvantaged than the populations served by ABC and PPP; and (iii) the quality offered at Head Start centers is heterogenous but on average is probably lower then the quality offered by ABC or PPP. Thus, it is not surprising that even after control contamination is taken into account, and a more clearly defined counterfactual identified, the estimated short-term impacts of Head Start are smaller than the impacts of the demonstration programs.

Long-run studies of Head Start based on observational data show substantial effects on later-life, socio-economic outcomes. These findings reinforce the need to consider multiple skills when evaluating early childhood programs. Dismissing Head Start as a failure because of a documented fadeout of IQ ignores the fact that early education has effects on multiple important dimensions of individual lifetimes. This is especially important because these dimensions may be complementary and self-productive. Negative assessments of Head Start ignore an important body of evidence.[66]

3.4.7 The Tennessee Voluntary Pre-Kindergarten Program

A recent evaluation of a means-tested local program in the US (The Tennessee Voluntary Pre-kindergarten Program) has recently captured public attention. This program is not a Head Start program. However, like Head Start, it is large-scale and targets children on the basis of socio-economic status. A handful of sites affiliated with the program are Head Start

[66] An illustrative example is Fox Business News (2014).

centers, although it is not clear whether any of these are included in the program's evaluation. This program is used as evidence against the effectiveness of large-scale preschool programs like Head Start (see Barshay, 2015). The Tennessee Voluntary Pre-kindergarten Program (TN-VPK) is a statewide kindergarten program, targeting disadvantaged 4 year-old children one year before kindergarten. It began as a pilot program in 1998 and became statewide in 2005. More details on its implementation, quality, and funding are reported in Appendix B.

The program is evaluated by a randomized control trial. However, the evaluation has major flaws and the interpretation of its results is clouded by the presence of control contamination. Program implementers requested parental consent *after* performing the randomization, causing substantial selective attrition from the study. The subsample for whom they received consent is called the Intensive Substudy. For the first cohort of participants, only 46% of the parents in the treatment group consented to enter the study and 32% of the parents in the control group consented. The rates of consent for the second cohort were 74% for the treatment group and 68% for the control group. This sampling plan creates a major problem of selective attrition. Experimental methods to evaluate this program become invalid, so the evaluators rely on non-experimental methods (Lipsey et al., 2013, 2015).[67]

The evaluation of TN-VPK does not account for control contamination. In their sample, 27% of the children in the control group attended Head Start or a private, center-based preschool program (Lipsey et al., 2015). The evaluation of this program does not address these confounds and does not identify a clear counterfactual.

A reduced set of measures were reported for the full sample, including grade repetition, attendance, disciplinary action, and special education. Estimates of these outcomes do not rely on flawed non-experimental methodology. The authors find that the treatment group was .77 percentage points less likely to repeat kindergarten. Short-term effects on cognition for the intensive substudy sample fade out or become negative as children age. The treat-

[67]To correct the selection problem caused by differential consent across control and treatment groups, the authors match on observable covariates. However, differential consent changed the composition of each group, and this methodology does not account for the resulting differences in unobserved characteristics.

ment group was 4 percentage points less likely to repeat a school grade. Short-term effects on cognition fade out. This evaluation does not represent strong evidence against the effectiveness of early childhood education programs. Instead, it illustrates that interpreting effects without accounting for flaws in experimental design or estimating clear counterfactuals produces misleading policy conclusions. It cautions against the use of randomized control trials as a gold standard. Evidence from non-experimental studies should not be outweighed by evidence from a randomized control trial without serious consideration of the methodologies of the individual studies.

3.5 Evidence from Large-Scale Programs

Evidence from demonstration programs and Head Start provides a strong case for the effectiveness of means-tested early childhood education in promoting child development. Moreover, the evidence from PPP and ABC shows that programs targeting disadvantaged children are socially and economically efficient. They also support work by mothers with young children. In this section, we study large-scale means-tested programs other than Head Start, and the evidence from universal programs.[68] Proposals have been made for universal programs (Office of the Mayor, New York City, 2014) and different forms of means-tested programs (The White House, 2014b).

The US government funds a variety of large-scale programs and initiatives. Table 3.10 describes the components of some major sources of federal funding for early childhood initiatives. There are two other major sources of funding: (i) Race to the Top: a source of funding for states, in which they compete on the basis of the quality, outcomes, and progress of their programs. States are selected for awards between 37.5 and 75 million 2014 USD (The White House, 2014b); and (ii) Preschool for All: an initiative providing 75 billion 2014 USD over ten years targeting low income ($\leq 200\%$ of the federal poverty line) 4 year-olds, with the aim

[68] A universal program is available to a general population of children in a local setting (e.g., county, state, country) when the only eligibility requirement is age.

of expanding the program to moderate-income children. Its goal is to increase the quality and quantity of available preschool and to support voluntary home visiting programs for the most disadvantaged families by providing grants to states to expand their existing preschool infrastructure and Head Start options (The White House, 2014b).

Though the evidence on preschool programs is limited by a dearth of non-cognitive and long-term measures, a clear pattern emerges. Universal programs are not universally effective. Results from several large-scale programs show that early childhood education is most effective when targeted toward disadvantaged children. Studies of childcare arrangements of children in the US indicate that impacts depend on the quality of the program being taken-up relative to the quality of the next best alternative. Because disadvantaged children typically have low-quality alternatives compared to advantaged children, they gain more from early childhood education.

The studies discussed in this section shed light on the potential benefits from universal programs and provide two major insights: (i) though they offer access with no eligibility constraints besides age, universal programs do not produce universal take-up; and (ii) disadvantaged children benefit the most from universal programs. This is a consequence of their having lower-quality alternatives compared to more advantaged children. There is also a hint that at current quality levels, universal programs may harm the children of affluent parents who have better alternatives. The magnitude of effects depends on the quality of the program relative to a child's alternative.[69]

The rest of this section proceeds as follows. First, we summarize studies of universal subsidies to childcare in Quebec, Canada and Norway (Section 3.5.1). Second, we summarize studies of a group of universal preschool programs in Oklahoma, Georgia, and Boston (Section 3.5.2). We then summarize the findings of the section (Section 3.5.3). We present detailed descriptions of these programs in Appendix B.

[69]Blau (2003) refers to center-based programs as formal programs and to non-center-based programs as informal programs. He notes that, generally, the quality of the former is higher than that of the latter. This section follows his characterization of childcare.

Table 3.10: Federal Funding Streams for Childcare

	Eligibility	Program Description	Program Requirements	Scope
Head Start, 1965-present	Children aged 3-5. Family income $\leq$ 190% Fed. income level.	Grants given to centers that provide development services, child care, parenting education, case management, health care (including referrals), nutrition, and family support. Can be Home-based (which includes weekly home visits and group socialization), center-based, family care, and mixed-approach.	Centers must follow curricular guidelines and pass teacher/staff qualification requirements and program quality and compliance evaluations.	2013 Federal Appropriation (including local projects and support activities): \$7.74 billion (2014 USD). 2013 Enrollment (including Migrant programs): 903,679.
Early Head Start, 1994-present	Expectant mothers and children under age 3. Family income $\leq$ 190% Fed. income level.	Grants given to centers that provide development services, child care, parenting education, case management, health care (including referrals), nutrition, and family support. Can be Home-based (which includes weekly home visits and group socialization), center-based, and mixed-approach.	Centers must follow curricular guidelines and pass teacher/staff qualification requirements and program quality and compliance evaluations.	2014 Federal Appropriation: \$1.37 billion (2014 USD). 2014 Enrollment: 115,826.
Child Care Development Fund (CCDF), 1990-present	Family income $\leq$ 85% of the state median income for a family of the same size. Children under 13.	Funds are granted to states that provide subsidies to families for the purpose of paying for childcare.	Few restrictions. Childcare facilities must meet state health/safety regulations. 2 % of funds must be allocated to educating families on childcare options.	2013 CCDF Federal-Only funding: \$5.10 billion (2014 USD). 2013 National "average monthly adjusted number of families and children served": 874,200 families and 1,455,100 children.
Individuals with Disabilities Education Act (IDEA) Preschool Grants, 1977-present	Preschool-aged (3-5) children who are experiencing developmental delays (as defined by state law) and need special education.	Funds are provided to states on the basis of the state's proportion of disabled children. They must be used on educational programs that promote school readiness and incorporate pre-literacy, language, and numeracy skills.	Children with disabilities must be educated with children who are not disabled.	2014 Federal allocations: \$353 million (2014 USD). 2014 Enrollment: 749,971 children.

Source: **HS and EHS** : Vogel et al. (2006), Love et al. (2002), Administration for Children and Families, Office of Head Start (2009). There are some exceptions to the income requirements for special needs children and certain minorities. Furthermore, up to 10% of enrollees in each center may have family income higher than the cutoff. **IDEA**: Administration for Children and Families, Office of Head Start (2014). **CCDF**: U.S. Department of Education (2015). Note: This table compares some of the major federal funding streams for public childcare. CCDF is also known as the Child Care and Development Block Grant (CCDBG). IDEA was passed in 1990 but was a continuation of the Education for All Handicapped Children Act, which was passed in 1975.

3.5.1 Universal Subsidies to Childcare

3.5.1.1 Norway

In 1975, the Norwegian parliament approved the Kindergarten Act, a reform which promoted a large-scale expansion of subsidized childcare. The reform was universal: all children from ages 3 to 6 were eligible, regardless of their family background. It led to a staged expansion inducing time and regional variation across 400 municipalities. The reform assigned responsibility for childcare provision to municipalities that followed federal quality standards, e.g., educational content, group size, staff skill composition, and physical environment. As a consequence of the reform, childcare coverage for children ages 3 to 6 increased from 10% in 1975 to 28% in 1979 (Havnes and Mogstad, 2011).[70]

Havnes and Mogstad (2011) exploit regional and time variation across municipalities in the roll-out of the reform to identify its effects using a standard difference-in-difference framework. They find positive effects of the program on a battery of long-term outcomes measured when participants were in their mid-30s, including years of education, college attendance, probability of being a high-school dropout, welfare dependency, and single parenthood.[71] They present two estimates. First, the intent-to-treat estimate, which simply compares eligible and ineligible children, given the time and regional variation. Second, they use a Bloom estimator to adjust the intent-to-treat estimate by the increase in childcare coverage.[72] In all cases, the effects are larger when adjusting for take-up. Applying the Bloom estimator produces a 7% increase in the probability of attending college, a 6% decrease in the probability of being a high school dropout, and a 5% decrease in the probability of being on welfare.

[70]The two main studies from which we draw results do not provide details on the characteristics of the families of children who used center-based care compared to those that did not. Thus, we cannot characterize the children who take-up the program and distinguish from those who did not. Drange et al. (2012) provide some related description of childcare take-up in Norway. As recently as 1996, relatively disadvantaged children under age 6 were under-represented in early childhood education participation.

[71]Examples of treatment effects include: an increase of .06 (s.e. .02) years of education; an increase of 1% (s.e. .3%) in college attendance; a decrease on the probability of being a dropout of 1% (s.e. .3%); and a decrease in welfare dependency of 1% (s.e. .3%).

[72]See Bloom (1984).

When they decompose results for a subsample of children of high school dropouts and high school graduates they find that the effects on education are driven primarily by children whose mothers are less educated. Estimates by gender show that females who received the treatment are less likely to be low earners and more likely to be average earners. This finding aligns with the evidence from ABC, indicating a positive treatment effect on age 30 income for women.

Although the authors do not explore the mechanisms driving their results, they provide a set of estimates that shed light on this. As discussed so far, they point out the relevance of considering children's next best alternative when the reform rolled out. They show that the reform had no effect on the amount of hours mothers work. However, it changes childcare take-up. The authors conclude that the reform crowds out informal childcare and increases the quality of the formal childcare taken up. Parents sent more children to center-based or formal childcare and less to informal care. Thus, the positive effects are a consequence of moving children from informal to formal care.

Havnes and Mogstad (2014) expand the analysis of Havnes and Mogstad (2011). They use the characteristics of the children who were affected by the reform and note that relatively disadvantaged children benefited the most from it. They allow for non-linearity in the differences-in-differences framework of Havnes and Mogstad (2011). Specifically, they explore variation in the effects of the reform on children along the earnings distribution once they become adults. They find that "upper-class children suffer a mean loss of $1.15 for every dollar spent on subsidized child care, whereas children of low-income parents experience an average gain of $1.31 for every dollar spent" (Havnes and Mogstad, 2014), which produces an increase in social mobility across the participating cohorts.

The evidence from this reform relates to two of the policy implications on which we present evidence throughout the paper. First, disadvantaged children benefit the most from early childhood education. In the case of Norway, it is very plausible that the reform crowded out poor informal alternatives for disadvantaged children, resulting in a relatively large

improvement in their early environments compared to those of advantaged children. This interpretation is further supported by the relatively larger effects for children of high school dropouts compared to children of high school graduates.

This point relates to the second implication. The quality of the early environments of children is fundamental. The reform in Norway made more slots available in formal or center-based care, which is relatively high-quality. This produces gains in short- and long-term outcomes for the neediest children.

3.5.1.2 Quebec

In 1997, the government of Quebec introduced a universal policy for families with children of ages 0 to 4. Regulated, center-based childcare was subsidized to have an effective price of at most 5.00 Canadian dollars[73] a day. All children aged 5 have access to free public kindergarten.[74]

Before 1997, only low-income families in Quebec received childcare subsidies. Further, low-income families ($\leq$ 57,680 2014 USD) received a 75% tax credit for childcare expenditures (Baker et al., 2005). This implies that the gain low-income families had from the 1997 reform was relatively small compared to the gain of high-income families. There are three components to the reform. First, for children younger than age 2, all previously informal childcare centers were certified and the staff was trained. Second, for children older than 2 but younger than kindergarten age, center-based childcare was subsidized. Third, kindergarten was made free.

Baker et al. (2008) evaluate the effects of the policy exploiting cross-Canada regional variation around the years of its implementation, comparing the pre- and post-policy outcomes of families in Quebec with the outcomes of families in the rest of Canada. They find that the effects of these reforms on child behavior and parent-child interactions are negative.

[73] 1997 dollars.

[74] Classroom size, caregiver education, and similar standards were imposed as part of the reform, one of its objectives being to improve the quality of childcare. More details are in Appendix B.

The policy caused a sizable increase in maternal labor supply (around 10 percentage points) with its effect mainly being experienced by high-income families, which the program dramatically changed the cost of childcare for. As a result, it crowded out parental care, which may be of a higher quality than center-based arrangements for some high-income families.

The policy increased emotional disorder and physical aggression at ages 2 and 3 and decreased social development at ages 0 to 3. Furthermore, it had negative effects on families in terms of effective parenting and maternal depression when children were between 0 and 4 years old.

Offsetting these negative findings, in later work, Baker et al. (2015) find that the policy had small, but beneficial effects for disadvantaged children. These include reduced hyperactivity, anxiety, and aggression at ages 2–3. Effects on non-cognitive outcomes are particularly strong for boys. Moreover, Baker et al. (2015) find evidence of decreased criminal activity as measured by apprehensions and convictions. The benefits reported in adolescence for disadvantaged boys is consistent with other evidence from programs targeted to disadvantaged families.

The 1997 reform in Quebec was implemented on top of existing subsidies to low-income families. It attracted more affluent families into the program by subsidizing childcare but not providing high-quality services at the level offered in affluent homes. The negative early-life results arise because: (i) disadvantaged families were already being offered a subsidy before the policy and centers for children above age 3 were certified and presumably high-quality; and (ii) the program crowded out maternal time spent on child care by relatively affluent families. This evidence underscores the importance, in any evaluation, of considering who took up the policy and what their next best alternative would have been in the absence of the policy.

3.5.2 Local Universal Programs in the US

For the universal public programs provided in Georgia and Oklahoma, some data on program take-up by socio-economic status are available. Universal access to programs does not imply universal take-up. In these programs, low socio-economic status is measured by eligibility for free or reduced price lunch, which requires that the child's family is at or below 185% of the federal poverty line. In Georgia, 59% of all preschool-age children in the state took up the program. Of these, 60% were eligible for free or reduced price lunch. In Oklahoma, 74% of all preschool-age children took up the program. Of these, 61% were eligible for free or reduced price lunch. Take-up is substantially lower among more affluent families.[75]

Cascio and Schanzenbach (2013) provide further evidence on take-up. By pooling data from Georgia and Oklahoma to make a comparison with the rest of the states in the US, they find that take-up differs across maternal education levels. Specifically, they find that between 4 and 5 out of every 10 children enrolled in public schools would have otherwise been enrolled in private preschools if their mothers had at least some college education. Thus, they project that the increase in preschool attendance in this relatively advantaged group is between 11 and 14 percentage points, compared to an increase of between 19 and 20 points for the pooled sample.

Georgia and Oklahoma sponsor preschool programs which have a relatively high score in the National Institute for Early Education Research (NIEER) quality index (Cascio and Schanzenbach, 2013), which is claimed to measure the quality of a state preschool program.[76] Georgia and Oklahoma have a high score because they require the teachers in every classroom to hold a bachelor's degree and have a certificate in early education. They also have class size

[75]Family poverty is defined in terms of family income starting below the 200% poverty line. Using elementary probability calculations and data on the percentage of children eligible for free or reduced price lunches (for which eligibility is determined by family income at or below the 185% poverty line), 49% of children in Oklahoma and Georgia were in poverty (American Community Survey) United States Census Bureau, 2014. Using the total take-up and take-up by socio-economic status statistics, the probability of taking-up the program for a child in a poor household is 79% in Georgia and 99% in Oklahoma. Similarly, the probability of taking-up the program for a child in a non-poor household is 40% in Georgia and 49% in Oklahoma.

[76]We note, however, that the Tennessee Program previously discussed also had a high NIEER quality index. See Lipsey et al., 2015. The validity of the NIEER score has not been established.

requirements—class size is capped at 20 children and a 1:10 teacher-student ratio is enforced. Both programs are partially funded through the Preschool for All initiative, though they also receive funding from other sources. Oklahoma's preschools are provided by public schools and they receive funding from state and federal sources. Though Georgia's preschools are publicly funded, the services are provided by private centers.

Cascio and Schanzenbach (2013) evaluate the Georgia and Oklahoma programs using a strategy similar to that of the evaluations of the Norway and Quebec reforms by exploiting regional and time variation across these and the rest of the states in the US. They estimate intent-to-treat effects of the policy on children up to eighth grade. Their findings indicate that disadvantaged children, as measured by their eligibility for free lunch, have substantial gains in reading and math test scores by fourth grade. The effects on reading vanish by eighth grade, but the effects on math scores remain statistically precise and are economically significant. For advantaged children, the effects become small by fourth grade and vanish by eighth grade. The authors present evidence on the mechanisms producing the effects. Disadvantaged children spend less time with their mothers, but the quality of the interaction increases because they spend more time reading, playing, and doing other activities together. That is, there is a relatively large improvement in the quality of the early environment for disadvantaged children.

The strategies used to identify the effects of the reforms in Norway and Quebec and the state programs in Georgia and Oklahoma are very similar. They exploit time and regional variation in program roll-out. In Norway, the reform was gradual and had time and regional variation across 400 municipalities. Thus, the estimates compare regions that differ in time of the policy implementation. In Quebec, the reform was introduced in the whole province and the estimates are identified by comparing outcomes in Quebec with those in the rest of Canada. Similarly, the state programs in the US are evaluated by comparing outcomes across Georgia and Oklahoma and the rest of the states in US.

There is a crucial drawback to this strategy, which is inherent in difference-in-difference

strategies. If there are any differences in trends of unobserved local characteristics across treatment and comparison group regions, then difference-in-difference estimates do not represent the effects of the reform, but rather differences in trends that would cause these effects even in the absence of the reform. In the example of Quebec, if previous policies uniquely changed the way in which the market for female labor increased in that province, and this caused the childcare decisions observed in the period after the reform, then the estimates of program effects on labor supply are contaminated by this pre-existing trend.

To assess this concern, in their study, Havnes and Mogstad (2011) perform a battery of robustness checks. These include different calculations of standard errors, such as clustering, to allow for various scenarios of unobserved correlation across municipalities, excluding cities from the sample, adding municipal fixed effects, and adding time trends interacted with multiple observed characteristics at municipality level. Their results are not sensitive to any of these sensitivity exercises. The fact that the reform in Norway was rolled out at municipality level provides a large amount of variation with which to perform many forms of sensitivity analyses.

Unfortunately, this is not the case for Quebec, as the reform was at the provincial level. Nevertheless, the authors of the Quebec study perform sensitivity analyses and report robust results. In the study of Cascio and Schanzenbach (2013), the authors perform sensitivity analysis by controlling for state trends and use a battery of observed characteristics. They also explore sensitivity with respect to the window of observations they consider. While these three studies differ in the degree to which they test for sensitivity, all find little evidence for it.

Gormley and Gayer (2005) and Gormley et al. (2005) evaluate Oklahoma's preschool program in a local setting. They use administrative data from Tulsa and exploit a sharp regression discontinuity design on age eligibility. Namely, children are eligible to attend preschool if they are 4 years of age by September 1st of the school year. Thus, they compare children of very similar ages who were just barely eligible with those who are just barely

ineligible. Data include tests measuring cognition for both groups. For the children who were not eligible, they use tests at preschool entry the following year. For the children who were eligible, they use tests at the end of preschool. They report a gain of 0.39 and 0.24 standard deviations in language and motor skills, respectively. However, this estimate is short-run in nature. The program accelerates academic competence but has no long-run effect. This evidence suggests that children in some form of schooling do better on tests than children not in school. After all children enter school, the effects vanish by grade 3 .[77]

Weiland and Yoshikawa (2013) evaluate a universal preschool program in Boston using a similar strategy. The program served 2,045 children in 69 elementary schools within the city. Any child turning 4 years-old before September 1st was eligible. Participants of the program received a year of free full-day pre-kindergarten in an urban public school. The children received a common curricula: full implementation of the literacy and language curriculum, Opening the World of Learning, and the mathematics curriculum, Building Blocks. Reports indicate that the curricula were implemented with high fidelity across preschools (Weiland and Yoshikawa, 2013).

The nature of the data makes it straightforward to compare children who were arbitrarily close to the eligibility cohort, but still not eligible, with those who were eligible and participated in the program. The reported results are positive on mathematics, reading, and some measures of social skills at the beginning of the first school year immediately following program completion. However, when they are disaggregated, these positive results show considerable variability. While children eligible for free lunch had impacts on self-control (0.3 effect size), ineligible children had no impacts on this dimension. Impacts in numeracy were very strong for both groups. The magnitudes of the effect sizes are .66 and .47, respectively.

We are skeptical about the interpretation of the estimates reported in Gormley and Gayer (2005), Gormley et al. (2005), and Weiland and Yoshikawa (2013). Their reported effects are short-run in nature and simply compare exposed children to unexposed children

[77]See Hill et al. (2012).

at the end of one year of the program. They do not account for catch-up in the scores when the unexposed children eventually enter school. Effects vanish by grade three in the Gormley studies. (Weiland and Yoshikawa, 2013 only analyze short-term outcomes measured in the fall after preschool completion.) An additional problem with these regression discontinuity studies is the large bandwidth often employed (i.e., a broad band of ages of children on which either side of the discontinuity point is used). There are few children available to identify the impact in the vicinity of the cutoff and there is selective attrition of children from samples.

3.5.3 Summary of the Evidence from Universal Programs

The evidence on universal programs supports a general finding consistent with the entire body of evidence in this paper. Disadvantaged children benefit more from early childcare education than do advantaged children. This is due to a larger improvement in the quality of the early environment for disadvantaged children compared to advantaged children. When children attend programs with higher quality care than they would have received at home or at an alternative setting, the effects of the programs are generally positive. Given that disadvantaged children have less access to alternatives, they benefit the most from universal programs. Programs that crowd out high-quality alternatives for advantaged children, as in Quebec, produce weak or even negative effects.

Further research is required to strengthen this body of evidence. In particular, the most rigorous analyses study policy changes and estimate their effects through reduced form estimates. Some of them shed light on the mechanisms driving the policy by exploring long-term effects, effects on maternal labor supply, etc. However, this literature could benefit from models that investigate the mechanisms through which estimated effects are generated.

3.6 The Importance of Quality

The studies discussed thus far indicate that when the childcare options for families are low in quality, center-based policies tend to have positive effects. This is especially true for disadvantaged families for whom alternatives are of relatively low quality. Following the recent literature, this section uses attendance to center-based care as an indicator for participation in a high-quality program and attendance to non-center-based care as an indicator for participation in a low-quality program. Generally speaking, center-based childcare establishments are required to be certified to be funded or run (see Appendix B). Disadvantaged children have less access to center-based childcare. All programs found to have positive effects have relatively high quality standards (see Appendices A and B). Blau and Currie (2006b) present an extensive survey of the market for childcare. They find that standards such as low staff-child ratios, small classroom size, and higher levels of teacher education contribute to the effectiveness of childcare centers.

Bernal (2008) and Bernal and Keane (2011) reinforce the evidence on the importance of quality by comparing the effects of center-based and non-center-based arrangements. They use the NLSY79 to examine childcare decisions in the US and their impacts on parental labor force participation and child development. They analyze the range of childcare options available in the US, including formal and informal care options. They use different methodologies to assess the impact of childcare on cognitive and non-cognitive development: (i) an approach using a fully structural model and (ii) an instrumental variables approach. The first paper uses a sample of married women. The second paper uses a sample of single mothers and exploits exogenous changes in welfare program structures as sources of variation affecting the probability of a child being in childcare. The papers show that childcare has negative effects on cognition at ages 5 to 8, with a magnitude of 0.13-0.14 standard deviations, and a standard error of .049. The negative effects arise from non-center-based childcare, while center-based childcare has no effect.

García et al. (2014) provide new insights using data from a demonstration program, IHDP. Using a methodology similar to that of Bernal and Keane (2011), but utilizing a more complete set of measures, they find that: (i) time spent with the mother and center-based childcare have positive effects that are very similar in magnitude on average; (ii) policies that give access to center-based childcare crowd out maternal time; and (iii) maternal time has strikingly different consequences for more or less disadvantaged children, reflecting the quality of home interactions; better home environments promote child development. Adverse home environments retard it.

3.7 Summary

Our analysis is based on three important principles from the literature on the economics of human development: (i) multiple skills beyond just cognition are important and are produced by effective programs; (ii) the skill formation process is dynamic and early home environments play a major role in shaping child lives; and (iii) answering policy questions requires consideration of the alternatives available to the targeted population.

Our main conclusion is that at current levels of quality provided, disadvantaged children benefit the most from early childhood education. The services offered improve on what is offered to them at home. The high-quality means-tested demonstration programs that we have examined are socially efficient as measured by benefit-cost ratios and rates of return. There is a strong case for high-quality means-tested early childhood education (using a broad definition of means-tested). The evidence for universal programs is somewhat ambiguous. The evidence from Quebec suggests that standard childcare programs supporting the market labor supply of affluent women may harm their children, but may aid the children of disadvantaged families.

These conclusions are based on the following bodies of evidence:

1. *From our primary analysis of the data on high quality demonstration programs, we*

conclude:

(a) Increases in cognition, as measured by IQ, generally fade out, but do not always disappear. However, gains in early life non-cognitive skills generate success later in life, boosting outcomes such as education, employment, health, and reduced criminal activity.

(b) Methodology is available to assess demonstration programs with compromised randomizations, small sample sizes, and attrition. Applying it shows that high quality demonstration programs have positive effects over the life-cycle. These effects survive conservative tests, adjusting test statistics for the effects of multiple hypotheses testing.

(c) When evaluated comprehensively, demonstration programs targeting disadvantaged populations are socially efficient, as measured by their rates of return and benefit-cost ratios.

2. *Head Start*

(a) Head Start provides heterogeneous treatment to heterogeneous populations. Therefore, when assessing its impacts, it is crucial for researchers to study the available alternatives in the settings where children take up treatment.

(b) Studies accounting for control group contamination—i.e., control group families that find alternative early childhood education environments outside the home—show that the short-run effects of Head Start on cognitive and non-cognitive skills are positive and moderate to strong.

(c) Studies evaluating long-term outcomes from Head Start find that the program has persistent beneficial effects on important later-life outcomes, such as health and education based on nationally representative data sets.

(d) Crude cost-benefit analyses of Head Start hint that the program might be socially efficient. More comprehensive evaluations likely imply high internal rates of returns, as current estimates only include gains in earnings.

3. *Universal Programs*

   Disadvantaged children benefit the most from universal programs offered at current quality levels. Advantaged children have enriched environments available to them and their parents are less likely to use them. In contrast, without access to such programs, disadvantaged children spend time in low-quality environments or informal settings.

3.8 Acknowledgements

This research was supported in part by the American Bar Foundation, the Pritzker Children's Initiative, the Buffett Early Childhood Fund, NIH grants NICHD R37HD065072, NICHD R01HD54702, and NIA R24AG048081, an anonymous funder, and Successful Pathways from School to Work, an initiative of the University of Chicago's Committee on Education funded by the Hymen Milgrom Supporting Organization. We are very grateful to Marianne Haramoto, Fernando Hoces, Joshua Ka Chun Shea, Matthew C. Tauzer, and Anna Ziff for research assistance and useful comments. We thank Robert Moffitt, David Blau, the other authors of this volume, and Raquel Bernal, Avi Feller, Micheal Keane, Patrick Kline, Sylvi Kuperman, and Rich Neimand for valuable comments. The views expressed in this paper are those of the authors and not necessarily those of the funders or persons named here or the official views of the National Institutes of Health.

Supplementary Files

The following appendices are available as supplementary files and online as follows:

Web Appendix: Chapter 1

The appendix can be found at `https://cehd.uchicago.edu/fadeout`

Web Appendix: Chapter 2

The appendix can be found at `https://cehd.uchicago.edu/ABC-CBA`

Web Appendix: Chapter 3

The appendix can be found at `https://cehd.uchicago.edu/ECE-US`

Bibliography

Administration for Children and Families, Office of Head Start (2009). Head Start Program performance standards and other regulations. Technical report, U.S. Department of Health and Human Services, Washington, DC.

Administration for Children and Families, Office of Head Start (2014). Head Start Program facts fiscal year 2014. Technical report, U.S. Department of Health and Human Services, Washington, DC.

Alison, M. (2014). How lasting are the benefits of preschool? *The Washington Post*.

Armor, D. (2014). We have no idea if universal preschool actually helps kids. *The Washington Post*.

Arrow, K. J. and D. Levhari (1969, September). Uniqueness of the internal rate of return with variable life of investment. *Economic Journal 79*(315), 560–566.

Bailey, D., G. Duncan, C. Odgers, and W. Yu (2015). Persistence and fadeout in the impacts of child and adolescent interventions. Technical report.

Bajaj, V. and S. Labaton (2009, February 1). Big risks for U.S. in trying to value bad bank assets. New York Times.

Baker, M., J. Gruber, and K. Milligan (2005, December). Universal childcare, maternal labor supply, and family well-being. Working Paper 11832, National Bureau of Economic Research.

Baker, M., J. Gruber, and K. Milligan (2008, August). Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy 116*(4), 709–745.

Baker, M., J. Gruber, and K. Milligan (2015, September). Non-cognitive deficits and young adult outcomes: The long-run impacts of a universal child care program. Working Paper 21571, National Bureau of Economic Research.

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children 5*(3), pp. 25–50.

Barnett, W. S. (2011). Effectiveness of Early Educational Intervention. *Science 333*(6045), 975–978.

Barnett, W. S. and L. N. Masse (2007, February). Comparative benefit-cost analysis of the Abecedarian program and its policy implications. *Economics of Education Review 26*(1), 113–125.

Barshay, J. (2015, October 5). Studies shed light on fleeting benefits of early childhood education. *U.S. News & World Report News*. Produced by The Hechinger Report.

Bartik, T. J., W. Gormley, and S. Adelstein (2012). Earnings benefits of Tulsa's pre-K program for different income groups. *Economics of Education Review 31*(6), 1143–1161.

Becker, G. S. (1991). *A Treatise on the Family*. Harvard university press.

Belfield, C. R., M. Nores, W. S. Barnett, and L. Schweinhart (2006). The HighScope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Followup. *Journal of Human Resources 41*(1), 162–190.

Bernal, R. (2008). The effect of maternal employment and child care on children's cognitive development. *International Economic Review 49*(4), 1173–1209.

Bernal, R. and M. P. Keane (2011, July). Child care choices and children's cognitive achievement: The case of single mothers. *Journal of Labor Economics 29*(3), 459–512.

Berrueta-Clement, J. R., L. J. Schweinhart, W. S. Barnett, A. S. Epstein, and D. P. Weikart (1984). *Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19*. Ypsilanti, MI: High/Scope Press.

Bertrand, M. and J. Pan (2011). The trouble with boys: Social influences and the gender gap in disruptive behavior. Working Paper 17541, National Bureau of Economic Research.

Bitler, M. P., H. W. Hoynes, and T. Domina (2014). Experimental evidence on distributional effects of Head Start. Working Paper 20434, National Bureau of Economic Research.

Blau, D. (2003). Child care subsidy programs. In R. A. Moffitt (Ed.), *Means-Tested Transfer Programs in the United States*, Chapter 7, pp. 443–516. Chicago: University of Chicago Press.

Blau, D. and J. Currie (2006a). Pre-school, Daycare, and After-school Care: Who's Minding the Kids? *Handbook of the Economics of Education 2*, 1163–1278.

Blau, D. and J. Currie (2006b). Preschool, daycare, and afterschool care: Who's minding the kids? In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Volume 2 of *Handbooks in Economics*, Chapter 20, pp. 1163–1278. Amsterdam: North-Holland.

Bloom, H. S. (1984, April). Accounting for no-shows in experimental evaluation designs. *Evaluation Review 8*(2), 225–246.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics 87*(1), 115–143.

Bond, T. N. and K. Lang (2013). The evolution of the black-white test score gap in grades k–3: The fragility of results. *Review of Economics and Statistics 95*(5), 1468–1479.

Borghans, L., H. Meijers, and B. ter Weel (2008, January). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry 46*(1), 2–12.

Brooks-Gunn, J., R. Gross, H. Kraemer, D. Spiker, and S. Shapiro (1992, June). Enhancing the cognitive outcomes of low birth weight, premature infants: For whom is the intervention most effective? *Pediatrics 89*(6, Part 2), 1209–1215.

Brooks-Gunn, J., F.-r. Liaw, and P. K. Klebanov (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics 120*(3), 350–359.

Brooks-Gunn, J., C. M. McCarton, P. H. Casey, M. C. McCormick, C. R. Bauer, J. C. Bernbaum, J. Tyson, M. Swanson, F. C. Bennett, D. T. Scott, et al. (1994). Early intervention in low-birth-weight premature infants: Results through age 5 years from the Infant Health and Development Program. *Journal of the American Medical Association 272*(16), 1257–1262.

Burchinal, M. R., F. A. Campbell, D. M. Bryant, B. H. Wasik, and C. T. Ramey (1997, October). Early intervention and mediating processes in cognitive performance of children of low-income African American families. *Child Development 68*(5), 935–954.

Burchinal, M. R., M. Lee, and C. T. Ramey (1989). Type of day-care and preschool intellectual development in disadvantaged children. *Child Development 60*(1), 128–137.

Bureau of Labor Statistics (2011). National longitudinal surveys: NLSY79 children and young adults. Website.

Bureau of Labor Statistics (2015). National longitudinal surveys: The NLSY79. Website.

Camilli, G., S. Vargas, S. Ryan, and W. S. Barnett (2010). Meta-analysis of the Effects of Early Education Interventions on Cognitive and Social Development. *The Teachers College Record 112*(3).

Campbell, F., G. Conti, J. J. Heckman, S. H. Moon, and R. Pinto (2014). The effects of early intervention on human development and social outcomes: Provisional evidence from abc and care. *The Economics Journal*, to appear. Unpublished manuscript, University of Chicago, Department of Economics.

Campbell, F. A., G. Conti, J. J. Heckman, S. H. Moon, R. Pinto, E. P. Pungello, and Y. Pan (2014). Early childhood investments substantially boost adult health. *Science 343*(6178), 1478–1485.

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

Campbell, F. A., C. T. Ramey, E. Pungello, J. Sparling, and S. Miller-Johnson (2002). Early childhood education: Young adult outcomes from the abecedarian project. *Applied Developmental Science 6*(1), 42–57.

Carneiro, P. and R. Ginja (2014, November). Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start. *American Economic Journal: Economic Policy 6*(4), 135–173.

Cascio, E. U. (2009). Do investments in universal early education pay off? long-term effects of introducing kindergartens into public schools. Technical report, National Bureau of Economic Research.

Cascio, E. U. and D. W. Schanzenbach (2013). The impacts of expanding access to high-quality preschool education. Working Paper 19735, National Bureau of Economic Research.

Cascio, E. U. and D. O. Staiger (2012). Knowledge, tests, and fadeout in educational interventions. Technical report, National Bureau of Economic Research.

Caucutt, E. M. and L. J. Lochner (2012). Early and late human capital investments, borrowing constraints, and the family. Working Paper 18493, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011, November). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics 126*(4), 1593–1660.

Claessens, A., M. Engel, and F. C. Curran (2013). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal*, 0002831213513634.

Clarke, S. H. and F. A. Campbell (1998). Can intervention early prevent crime later? The Abecedarian Project compared with other programs. *Early Childhood Research Quarterly 13*(2), 319–343.

Conti, G., J. J. Heckman, and R. Pinto (2015). The long-term health effects of early childhood interventions. Forthcoming, *Economic Journal*.

Cunha, F. (2015). Subjective rationality, parenting styles, and investments in children. In P. R. Amato, A. Booth, S. M. McHale, and J. Van Hook (Eds.), *Families in an Era of Increasing Inequality: Diverging Destinies*, National Symposium on Family Issues Series, Chapter 6, pp. 83–94. New York: Springer.

Cunha, F., I. T. Elo, and J. Culhane (2013, June). Eliciting maternal expectations about the technology of cognitive skill formation. Working Paper 19144, National Bureau of Economic Research.

Cunha, F. and J. J. Heckman (2007, May). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Cunha, F. and J. J. Heckman (2008a). Formulating, Identifying, and Estimating the Technology of Cognitive and Non-cognitive Skill Formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F. and J. J. Heckman (2008b, Fall). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F. and J. J. Heckman (2009, April–May). The economics and psychology of inequality and human development. *Journal of the European Economic Association 7*(2–3), 320–364.

Cunha, F., J. J. Heckman, and S. Navarro (2005, April). Separating uncertainty from heterogeneity in life cycle earnings, The 2004 Hicks Lecture. *Oxford Economic Papers 57*(2), 191–261.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010a). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010b, May). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Currie, J. (2001a). Early Childhood Education Programs. *Journal of Economic Perspectives*, 213–238.

Currie, J. (2001b, Spring). Early childhood education programs. *Journal of Economic Perspectives 15*(2), 213–238.

Currie, J. and D. Thomas (1995a, June). Does Head Start Make a Difference? *American Economic Review 85*(3), 341–64.

Currie, J. and D. Thomas (1995b, June). Does Head Start make a difference? *American Economic Review 85*(3), 341–364.

Currie, J. and D. Thomas (2000). School quality and the longer-term effects of head start. *Journal of Human Resources 35*(4), 755–74.

Dalmia, S. and L. Snell (2008). Protect our kids from preschool. *The Wall Street Journal*.

Deming, D. (2009, July). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics 1*(3), 111–134.

Drange, N., T. Havnes, and A. M. J. Sandsør (2012, November). Kindergarten for all: Long run effects of a universal intervention. Discussion Paper 6986, Institute for the Study of Labor.

Duncan, G. J. and K. Magnuson (2013a). Investing in preschool programs. *The Journal of Economic Perspectives 27*(2), 109–132.

Duncan, G. J. and K. Magnuson (2013b). Investing in preschool programs. *Journal of Economic Perspectives 27*(2), 109–132.

Duncan, G. J. and R. J. Murnane (Eds.) (2011). *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*. New York: Russell Sage Foundation.

Duncan, G. J. and R. J. Murnane (2014). *Restoring Opportunity: The Crisis of Inequality and the Challenge for American Education*. Cambridge, MA/New York: Harvard Education Press/Russell Sage Foundation.

Duncan, G. J. and A. J. Sojourner (2013a). Can intensive early childhood intervention programs eliminate income-based cognitive and achievement gaps? *Journal of Human Resources 48*(4), 945–968.

Duncan, G. J. and A. J. Sojourner (2013b). Can intensive early childhood intervention programs eliminate income-based cognitive and achievement gaps? *Journal of Human Resources 48*(4), 945–968.

Eckenrode, J., M. Campa, D. W. Luckey, C. R. Henderson, R. Cole, H. Kitzman, E. Anson, K. Sidora-Arcoleo, and D. L. Olds (2010, January). Long-term effects of prenatal and infancy nurse home visitation on the life course of youths: 19-year follow-up of a randomized trial. *Journal of the American Medical Association 164*(1), 9–15.

Elango, S., A. Hojman, J. L. García, and J. J. Heckman (2015). Early childhood education. Forthcoming, in Moffitt, Robert (ed.), *Means-tested Transfer Programs in the United States II. Chicago: University of Chicago Press, 2016.*

Elango, S., A. Hojman, J. L. García, and J. J. Heckman (2016). Early childhood education. Forthcoming, in Moffitt, Robert (ed.), *Means-Tested Transfer Programs in the United States II*. Chicago: University of Chicago Press, 2016.

Engel, M., A. Claessens, and M. A. Finch (2013). Teaching students what they already know? the (mis) alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis 35*(2), 157–178.

Feller, A., T. Grindal, L. Miratrix, and L. Page (2016). Compared to what? Variation in the impacts of early childhood education by alternative care-type settings. Forthcoming, *Annals of Applied Statistics*.

Fox Business News (2014). Head Start has little effect by grade school? Video.

Frank Porter Graham Child Development Center (1979). The Frank Porter Graham Child Development Center progress report 1979. Technical report, Child Development Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Garber, H. L. (1988). *The Milwaukee Project: Preventing Mental Retardation in Children at Risk*. Washington, DC: American Association on Mental Retardation.

Garces, E., D. Thomas, and J. Currie (2002, September). Longer-term effects of Head Start. *American Economic Review 92*(4), 999–1012.

García, J. L. (2014). Ability, character, and social mobility. University of Chicago, Department of Economics.

García, J. L. (2015). Childcare and parental investment: Short and long-term effects. University of Chicago, Department of Economics.

García, J. L., J. J. Heckman, A. 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. University of Chicago, Department of Economics.

García, J. L., J. J. Heckman, A. Hojman, D. E. Leaf, M. J. Prados, J. Shea, and J. C. Torcasso (2016). Analyzing the Short- and Long-term Effects of Early Childhood Education on Multiple Dimensions of Human Development.

García, J. L., A. Hojman, and J. Shea (2014). The opportunity cost of early childhood education: Formal, informal and maternal care. University of Chicago, Department of Economics.

Gelber, A. and A. Isen (2013). Children's schooling and parents' behavior: Evidence from the head start impact study. *Journal of Public Economics*.

Gertler, P., J. J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. Chang, and S. M. Grantham-McGregor (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science 344*(6187), 998–1001.

Gilhousen, M. R., L. F. Allen, L. M. Lasater, D. M. Farrell, and C. R. Reynolds (1990). Veracity and vicissitude: A critical look at the Milwaukee Project. *Journal of School Psychology 28*(4), 285–299.

Goldman, D. P., D. Lakdawalla, P.-C. Michaud, C. Eibner, Y. Zheng, A. Gailey, I. Vaynman, J. Sullivan, B. Tysinger, and D. Ermini Leaf (2015). The Future Elderly Model: Technical documentation. Technical report, University of Southern California.

Gormley, Jr., W. T. and T. Gayer (2005). Promoting school readiness in Oklahoma: an evaluation of Tulsa's pre-K program. *Journal of Human Resources 40*(3), 533–558.

Gormley, Jr., W. T., T. Gayer, D. Phillips, and B. Dawson (2005, November). The effects of universal pre-K on cognitive development. *Developmental Psychology 41*(6), 872–884.

Gray, S. W. and R. A. Klaus (1970). The Early Training Project: a seventh-year report. *Child Development*, 909–924.

Gray, S. W., B. K. Ramsey, and R. A. Klaus (1982a). *From 3 to 20: The Early Training Project.* Baltimore: University Park Press.

Gray, S. W., B. K. Ramsey, and R. A. Klaus (1982b). *From 3 to 20: The Early Training Project.* Baltimore, MD: University Park Press.

Gross, R. T., D. Spiker, N. A. Constantine, W. L. Kreitman, C. W. Haynes, C. T. Ramey, D. Bryant, J. Sparling, B. H. Wasik, I. Lewis, et al. (1990). Enhancing the outcomes of low-birth-weight, premature infants. *Journal of the American Medical Association 263*(22), 3035–3042.

Gross, R. T., D. Spiker, and C. W. Haynes (1997). *Helping Low Birth Weight, Premature Babies: The Infant Health and Development Program*. Stanford, CA: Stanford University Press.

Guernsey, L. and L. Bornfreund (2013). Why preschool isn't enough. *The Atlantic Magazine*.

Havnes, T. and M. Mogstad (2011, May). No child left behind: Subsidized child care and children's long-run outcomes. *American Economic Journal: Economic Policy 3*(2), 97–129.

Havnes, T. and M. Mogstad (2014). Is universal child care leveling the playing field? *Journal of Public Economics 127*, 100–114.

Heckman, J., S. H. Moon, R. Pinto, P. Savelyev, and A. Yavitz (2010a). Analyzing Social Experiments as Implemented: a Reexamination of the Evidence from the HighScope Perry Preschool Program. *Quantitative Economics 1*(1), 1–46.

Heckman, J. J. (1992). Randomization and social policy evaluation. In C. Manski and I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs*, pp. 201–230. Cambridge, MA: Harvard University Press.

Heckman, J. J. (2008, July). Schools, skills and synapses. *Economic Inquiry 46*(3), 289–324.

Heckman, J. J. (2015, October). Analyzing the impacts of two influential early childhood programs on participants through midlife. Proposal submitted to the National Institutes of Health on October 5, 2015.

Heckman, J. J., N. Hohmann, J. Smith, and M. Khoo (2000, May). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics 115*(2), 651–694.

Heckman, J. J., A. Hojman, and J. C. Torcasso (2014). Forecasting the long-term effectiveness of early childhood interventions: The case of Head Start. University of Chicago, Department of Economics.

Heckman, J. J., M. Holland, T. Oey, D. L. Olds, R. Pinto, and M. Rosales (2014). A reanalysis of the Nurse Family Partnership Program: The Memphis randomized control trial. University of Chicago, Department of Economics.

Heckman, J. J., J. E. Humphries, and T. Kautz (Eds.) (2014). *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. Chicago: University of Chicago Press.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997, October). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies 64*(4), 605–654. Special Issue: Evaluation of Training and Other Social Programmes.

Heckman, J. J., H. Ichimura, and P. E. Todd (1998, April). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*(2), 261–294.

Heckman, J. J. and T. Kautz (2012a). Hard evidence on soft skills. *Labour economics 19*(4), 451–464.

Heckman, J. J. and T. Kautz (2012b, August). Hard evidence on soft skills. *Labour Economics 19*(4), 451–464. Adam Smith Lecture.

Heckman, J. J. and T. Kautz (2014). Fostering and measuring skills: Interventions that improve character and cognition. In J. J. Heckman, J. E. Humphries, and T. Kautz (Eds.), *The Myth of Achievement Tests: The GED and the Role of Character in American Life*, Chapter 9, pp. 341–430. Chicago: University of Chicago Press.

Heckman, J. J., S. Kuperman, and C. Cheng (2015). Understanding and comparing the mechanisms producing the impacts of major early childhood programs with long-term follow-up. University of Chicago, Department of Economics.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010b). The Rate of Return to the HighScope Perry Preschool Program. *Journal of Public Economics 94*(1), 114–128.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010c, August). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics 1*(1), 1–46.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010d, February). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics 94*(1–2), 114–128.

Heckman, J. J. and S. Mosso (2014a). The economics of human development and social mobility. *Annual Review of Economics 6*(1), 689–733.

Heckman, J. J. and S. Mosso (2014b, August). The economics of human development and social mobility. *Annual Review of Economics 6*(1), 689–733.

Heckman, J. J. and R. Pinto (2015). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric Reviews 34*(1–2), 6–31.

Heckman, J. J., R. Pinto, and P. A. Savelyev (2013a). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *Forthcoming in American Economic Review*.

Heckman, J. J., R. Pinto, and P. A. Savelyev (2013b). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–2086.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144. Amsterdam: Elsevier.

Hill, C. J., W. T. Gormley, Jr., and S. Adelstein (2012). Do the short-term effects of a strong preschool program persist? Working paper, Center for Research on Children in the U.S.

Hojman, A. (2015). Evidence on the fade-out of IQ gains from early childhood interventions: A skill formation perspective. University of Chicago, Center for the Economics of Human Development.

Iverson, G. and J.-C. Falmagne (1985). Statistical issues in measurement. *Mathematical Social Sciences 10*(2), 131–153.

Jacob, B. and J. Rothstein (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives*.

Kagitcibasi, C., D. Sunar, S. Bekman, N. Baydar, and Z. Cemalcilar (2009). Continuing effects of early enrichment in adult life: The Turkish Early Enrichment Project 22 years later. *Journal of Applied Developmental Psychology 30*(6), 764–779.

Kerr, M. A., R. E. Tremblay, L. Pagani, and F. Vitaro (1997). Boys' behavioral inhibition and the risk of later delinquency. *Archives of General Psychiatry 54*(9), 809–816.

Kirp, D. (2015). Does pre-k make any difference? *The New York Times*.

Klaus, R. A. and S. W. Gray (1968). The Early Training Project for disadvantaged children: A report after five years. *Monographs of the Society for Research in Child Development*, iii–66.

Kline, P. and M. Tartari (2015). Bounding the labor supply responses to a randomized welfare experiment: A revealed preference approach. Revise and resubmit, *American Economic Review*.

Kline, P. and C. Walters (2014, December). Evaluating public programs with close substitutes: The case of Head Start. IRLE Working Paper 123-14, Institute for Research on Labor and Employment, University of California–Berkeley.

Kline, P. and C. Walters (2015). Evaluating public programs with close substitutes: The case of Head Start. Working Paper 21658, National Bureau of Economic Research.

Knudsen, E. I., J. J. Heckman, J. Cameron, and J. P. Shonkoff (2006, July). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences 103*(27), 10155–10162.

Krantz, D., D. Luce, P. Suppes, and A. Tversky (1971). Foundations of measurement, vol. i: Additive and polynomial representations.

Kuperman, S. (2014). Interviews of Louise Derman-Sparks and Evelyn Moore (PPP teachers). Technical report, University of Chicagos Center for the Economics of Human Development, Chicago, IL.

Kuperman, S. (2014-2015). Interviews of Frances Campbell, Carrie Bynum, Phyllis Royster, Gael McGinness, Joseph Sparling, Albert Collier, Barbara Wasik, Lynne Vernon-Feagans, Tom Richey, Margaret Burchinal, Thelma Harms, and Richard Clifford. University of Chicago's Center for the Economics of Human Development, Chicago.

Kuperman, S. and C. Cheng (2014). Interviews of Frances Campbell, Carrie Bynum, Phyllis Royster, Gael McGinness, Joseph Sparling, Albert Collier, Barbara Wasik, Lynne Vernon-Feagans, Tom Richey, Margaret Burchinal, Thelma Harms, Richard clifford, Ron Haskins, and Susann Hutaff Haskins. University of Chicago's Center for the Economics of Human Development, Chicago.

Lavigueur, S., R. E. Tremblay, and J.-F. Saucier (1995). Interactional processes in families with disruptive boys: Patterns of direct and indirect influence. *Journal of Abnormal Child Psychology 23*(3), 359–378.

Lazar, I., R. Darlington, H. Murray, J. Royce, A. Snipper, and C. T. Ramey (1982). Lasting effects of early education: A report from the consortium for longitudinal studies. *Monographs of the Society for Research in Child Development*, i–151.

Leak, J., G. J. Duncan, W. Li, K. Magnuson, H. Schindler, and H. Yoshikawa (2010). Is timing everything? how early childhood education program impacts vary by starting age, program duration and time since the end of the program.

Lee, V. E. and S. Loeb (1995). Where do head start attendees end up? one reason why preschool effects fade out. *Educational evaluation and policy analysis 17*(1), 62–82.

Lehmann, E. and J. P. Romano (2005, June). Generalizations of the familywise error rate. *Annals of Statistics 33*(3), 1138–1154.

Lipsey, M. W., D. C. Farran, and K. G. Hofer (2015). A randomized control trial of the effects of a statewide voluntary prekindergarten program on children's skills and behaviors through third grade. Research report, Peabody Research Institute, Vanderbilt University, Nashville, TN.

Lipsey, M. W., K. G. Hofer, N. Dong, D. C. Farran, and C. Bilbrey (2013). Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and first grade follow-up results from the randomized control design. Research report, Peabody Research Institute, Vanderbilt University, Nashville, TN.

Love, J. M., E. Eliason Kisker, C. Ross, H. Raikes, J. Constantine, K. Boller, R. Chazen-Cohen, J. Brooks-Gunn, L. B. Tarullo, C. Brady-Smith, A. Sidle Fuligni, P. Z. Schochet,

D. Paulsell, and C. Vogel (2005). The effectiveness of early Head Start for 3-year-old children and their parents: Lessons for policy and programs. *Developmental Psychology 41*(6), 885–901.

Love, J. M., E. E. Kisker, C. M. Ross, P. Z. Schochet, J. Brooks-Gunn, D. Paulsell, K. Boller, J. Constantine, C. Vogel, A. S. Fuligni, and C. Brady-Smith (2002, June). Making a difference in the lives of infants and toddlers and their families: The impacts of early Head Start. Volumes I-III: Final technical report and appendixes and local contributions to understanding the programs and their impacts. Technical Report ED472186, Mathematica Policy Research.

Luce, R. D. and L. Narens (2008). *measurement, theory of*. Palgrave Macmillan.

Ludwig, J. and D. L. Miller (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity approach. *Quarterly Journal of Economics 122*(1), 159–208.

Ludwig, J. and D. A. Phillips (2008). Long-term effects of Head Start on low-income children. *Annals of the New York Academy of Sciences 1136*(1), 257–268.

Magnuson, K., C. Ruhm, and J. Waldfogel (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly 22*, 18–38.

Martin, A., J. Brooks-Gunn, P. Klebanov, S. L. Buka, and M. C. McCormick (2008). Long-term maternal effects of early childhood intervention: Findings from the infant health and development program (ihdp). *Journal of Applied Developmental Psychology 29*(2), 101–117.

Mâsse, L. C. and R. E. Tremblay (1997). Behavior of boys in kindergarten and the onset of substance use during adolescence. *Archives of General Psychiatry 54*(1), 62–68.

Masse, L. N. and W. S. Barnett (2002). *A Benefit Cost Analysis of the Abecedarian Early Childhood Intervention*. New Brunswick, NJ: National Institute for Early Education Research.

Mayer, S. E. (1997). *What Money Can't Buy: Family Income and Children's Life Chances*. Cambridge, MA: Harvard University Press.

McCollister, K. E., M. T. French, and H. Fang (2010). The cost of crime to society: New crime-specific estimates for policy and program evaluation. *Drug and Alcohol Dependence 108*(1–2), 98–109.

McCormick, M. C., J. Brooks-Gunn, S. L. Buka, J. Goldman, J. Yu, M. Salganik, D. T. Scott, F. C. Bennett, L. L. Kay, J. C. Bernbaum, C. R. Bauer, C. Martin, E. R. Woods, A. Martin, and P. H. Casey (2006, March). Early intervention in low birth weight premature infants: Results at 18 years of age for the Infant Health and Development Program. *Pediatrics 117*(3), 771–780.

McKey, R. H., L. Condelli, H. Ganson, B. J. Barrett, and C. McConkey (1985, June). The impact of Head Start on children, families and communities: Final report of the Head Start evaluation, synthesis and utilization project. Technical report, U.S. Department of Health and Human Services, Washington, DC.

McLanahan, S. (2004, November). Diverging destinies: How children are faring under the second demographic transition. *Demography 41*(4), 607–627.

McLanahan, S. and C. Percheski (2008, August). Family structure and the reproduction of inequalities. *Annual Review of Sociology 34*(1), 257–276.

Nagin, D. S. and R. E. Tremblay (2001, April). Parental and early childhood predictors of persistent physical aggression in boys from kindergarten to high school. *Archives of General Psychiatry 58*(4), 389–394.

Narens, L. (1981). On the scales of measurement. *Journal of Mathematical Psychology 24*(3), 249–275.

Noll, S. and J. Trent (Eds.) (2004). *Mental Retardation in America: A Historical Reader (The History of Disability)*. New York: NYU Press.

Office of the Mayor, New York City (2014). Ready to launch: New York City's implementation plan for free, high-quality, full-day universal pre-Kindergarten. Technical report, New York Department of Education.

Ohio University and Westinghouse Learning Corporation (1969). The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development (executive summary). Technical report, Ohio University and Westinghouse Learning Corporation, Athens, OH and New York.

Olds, D. L. (2006). The Nurse-Family Partnership: An evidence-based preventive intervention. *Infant Mental Health Journal 27*(1), 5–25.

Olds, D. L., C. R. Henderson, R. Chamberlin, and R. Tatelbaum (1986). Preventing child abuse and neglect: A randomized trial of nurse home visitation. *Pediatrics 78*(1), 65–78.

Olds, D. L., C. R. Henderson, and H. Kitzman (1994). Does prenatal and infancy nurse home visitation have enduring effects on qualities of parental caregiving and child health at 25 to 50 months of life? *Pediatrics 93*(1), 89–98.

Page, E. B. (1972). Miracle in Milwaukee: Raising the IQ. *Educational Researcher 1*(10), 8–10,15–16.

Panel Study of Income Dynamics (2015). PSID: A national study of socioeconomics and health over lifetimes and across generations. Website.

Parker, E., B. Atchison, and E. Workman (2016). 50-state review: State pre-K funding for 2015-16 fiscal year: National trends in state preschool funding. Technical report, Education Commission of the States, Denver, CO.

Project Head Start (1969). *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development*. Bladensburg, MD: Westinghouse Learning Corporation.

Puma, M., S. Bell, R. Cook, and C. Heid (2012). Head Start Impact Study: Final report. Technical report, Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC.

Puma, M., S. Bell, R. Cook, C. Heid, P. Broene, F. Jenkins, A. Mashburn, and J. Downer (2012). Third grade follow-up to the Head Start Impact Study: Final report. OPRE Report 2012–45, Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC.

Puma, M., S. Bell, R. Cook, C. Heid, and M. Lopez (2005, June). Head Start Impact Study: First year findings. Technical report, Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC.

Putnam, R. D. (2015). *Our Kids: The American Dream in Crisis*. New York: Simon and Schuster.

Raine, A., J. Liu, P. H. Venables, S. A. Mednick, and C. Dalais (2010). Cohort profile: The Mauritius Child Health Project. *International Journal of Epidemiology 39*(6), 1441–1451.

Ramey, C. T., G. D. McGinness, L. Cross, A. M. Collier, and S. Barrie-Blackley (1982). The Abecedarian approach to social competence: Cognitive and linguistic intervention for disadvantaged preschoolers. In K. M. Borman (Ed.), *The Social Life of Children in a Changing Society*, Chapter 7, pp. 145–174. Hillsdale, NJ: Lawrence Erlbaum Associates.

Reynolds, A. J. and J. A. Temple (1998, February). Extended early childhood intervention and school achievement: Age 13 findings from the Chicago Longitudinal Study. *Child Development 69*(1), 231–246.

Reynolds, A. J. and J. A. Temple (2006). Economic returns of investments in preschool education. In E. F. Zigler, W. S. Gilliam, and S. S. Jones (Eds.), *A Vision For Universal Preschool Education*, pp. 37–68. New York: Cambridge University Press.

Reynolds, A. J., J. A. Temple, B. A. B. White, S.-R. Ou, and D. L. Robertson (2011, January–February). Age 26 cost-benefit analysis of the Child-Parent Center early education program. *Child Development 82*(1), 379–404.

Ricciuti, A. E., R. G. St. Pierre, W. Lee, and A. Parsad (2004). Third national Even Start evaluation: Follow-up findings from the experimental design study. Technical report, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Washington, DC.

Romano, J. P. and A. M. Shaikh (2006, August). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics 34*(4), 1850–1873.

Romano, J. P., A. M. Shaikh, and M. Wolf (2010). Multiple testing. *The New Palgrave Dictionary of Economics*. Forthcoming.

Schneider, B. and S.-K. McDonald (Eds.) (2006). *Scale-Up in Education*, Volume 2: Issues in Practice. Blue Ridge Summit, PA: Rowman & Littlefield Publishers.

Smith, L. K. (2016, February). President's early learning budget for FY2017 and legislation to strengthen child care for families with young children. Website, Administration for Children and Families, http://www.acf.hhs.gov/blog/2016/02/presidents-2017-early-learning-budget (Accessed 4/17/16).

Snyder, T., C. de Brey, and S. Dillow (2016). *Digest of education statistics 2014*. National Center for Education Statistics.

Sommer, R. and B. A. Sommer (1983). Mystery in Milwaukee: Early intervention, IQ, and psychology textbooks. *American Psychologist 38*(9), 982–85.

St. Pierre, R. G., B. D. Layzer, L. Goodson, and L. S. Bernstein (1999). The effectiveness of comprehensive case management interventions: Evidence from the national evaluation of the Comprehensive Child Development Program. *American Journal of Evaluation 20*(1), 15–34.

St. Pierre, R. G., J. I. Layzer, B. D. Goodson, and L. S. Bernstein (1997). National impact evaluation of the Comprehensive Child Development Program: Final report. Technical report, Abt Associates Inc., Cambridge, MA.

The White House (2014a). The economics of early childhood investments. Technical report, Executive Office of the President of the United States, Washington, DC.

The White House (2014b). Fact sheet: Invest in US: The White House Summit on Early Childhood Education. Technical report, Office of the Press Secretary, Washington, DC.

Todd, P. E. and K. I. Wolpin (2003, February). On the specification and estimation of the production function for cognitive achievement. *Economic Journal 113*(485), F3–33.

Todd, P. E. and K. I. Wolpin (2007, Winter). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital 1*(1), 91–136.

Trahan, L. H., K. K. Stuebing, J. M. Fletcher, and M. Hiscock (2014). The flynn effect: A meta-analysis. *Psychological Bulletin 140*(5), 1332.

United States Census Bureau (2014). American Community Survey. Data set, United States Census Bureau.

U.S. Department of Education (2015). Preschool grants for children with disabilities: Funding status. Website.

U.S. Office of Management and Budget (2015). *Fiscal Year 2016 Budget of the U.S. Government*. Washington, DC: U.S. Government Printing Office.

Various (2014). Early childhood education for low-income students: A review of the evidence and benefit-cost analysis. Technical report, Washington State Institute for Public Policy.

Vogel, C. A., N. Aikens, A. Burwick, L. Hawkinson, A. Richardson, L. Mendenko, and R. Chazan-Cohen (2006, December). Findings from the survey of early Head Start programs: Communities, programs, and families. Final report. Technical Report ED498072, U.S. Department of Health and Human Services.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Wasik, B. H., C. Ramey, D. M. Bryant, and J. J. Sparling (1990, December). A longitudinal study of two early intervention strategies: Project CARE. *Child Development 61*(6), 1682–1696.

Weikart, D. P. (1967). Preliminary results from a longitudinal study of disadvantaged preschool children. Unpublished manuscript, ERIC No. ED 030 490. Presented at the 1967 convention of the Council for Exceptional Children, St. Louis, MO.

Weikart, D. P. (1970). *Longitudinal Results of the Ypsilanti Perry Preschool Project*, Volume 1 of *Monographs of the High/Scope Educational Research Foundation*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Weikart, D. P. et al. (1967). Preschool Intervention –A Preliminary Report of the Perry Preschool Project.

Weikart, D. P., J. T. Bond, and J. T. McNeil (1978). *The Ypsilanti Perry Preschool Project: Preschool years and longitudinal results through fourth grade*. High/Scope Educational Research Foundation.

Weiland, C. and H. Yoshikawa (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development 84*(6), 2112–2130.

White, J. L., T. E. Moffitt, A. Caspi, D. J. Bartusch, D. J. Needles, and M. Stouthamer-Loeber (1994). Measuring impulsivity and examining its relationship to delinquency. *Journal of Abnormal Psychology 103*(2), 192–205.

Yi, J., J. J. Heckman, J. Zhang, and G. Conti (2015, November). Early health shocks, intra-household resource allocation and child outcomes. *Economic Journal 125*(588), F347–F371.

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

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

Zhai, F. H., J. Brooks-Gunn, and J. Waldfogel (2014, December). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology 50*(12), 2572–2586.

Zigler, E. and S. Muenchow (1994). *Head Start: The Inside Story Of America's Most Successful Educational Experiment*. New York: Basic Books.