NBER WORKING PAPER SERIES

EARLY CHILDHOOD EDUCATION

Sneha Elango
Jorge Luis García
James J. Heckman
Andrés Hojman

Working Paper 21766
http://www.nber.org/papers/w21766

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2015

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 chapter 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 or of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Early Childhood Education
Sneha Elango, Jorge Luis García, James J. Heckman, and Andrés Hojman
NBER Working Paper No. 21766
November 2015, December 2015
JEL No. C93,I28,J13

ABSTRACT

This paper organizes and synthesizes the literature on early childhood education and childcare. In it, we go beyond meta-analysis and reanalyze primary data sources in a common framework. We consider the evidence from means-tested demonstration programs, large-scale means-tested programs and universal programs without means testing. We discuss which programs are beneficial and whether they are cost-effective for certain populations. The evidence from high-quality demonstration programs targeted toward disadvantaged children shows beneficial effects. Returns exceed costs, even accounting for the deadweight loss of collecting taxes. When proper policy counterfactuals are constructed, Head Start has beneficial effects on disadvantaged children compared to home alternatives. Universal programs benefit disadvantaged children.

Sneha Elango
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago IL 60637
selango@uchicago.edu

Jorge Luis García
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago IL 60637
jorge.ggmenendez@gmail.com

James J. Heckman
Department of Economics
The University of Chicago
1126 E. 59th Street
Chicago, IL 60637
and IZA
and also NBER
jjh@uchicago.edu

Andrés Hojman
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago IL 60637
andreshojman@gmail.com

Supplementary materials available at http://www.nber.org/papers/w21766:
- Web appendix

Contents

List of Figures

List of Tables

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); Reardon et al. (2011).
[2] McLanahan (2004); Heckman (2008).
[3] McLanahan and Percheski (2008).
[4] Putnam (2015).
[5] Office of the Mayor (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 the economic benefits of these programs exceed their costs.

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 (2006), Duncan and Magnuson (2013), and Yoshikawa et al. (2013). 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. (2014); 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 1 summarizes the programs we discuss and their basic features. We present detailed descriptions of these programs in Sections 3–5 and Appendices A and B. Section 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 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 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 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 6 discusses non-experimental evidence on the importance of quality environments in promoting child development. We summarize our findings in Section 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.

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 1: Comparing Demonstration Programs, Head Start, and Universal Preschool Programs

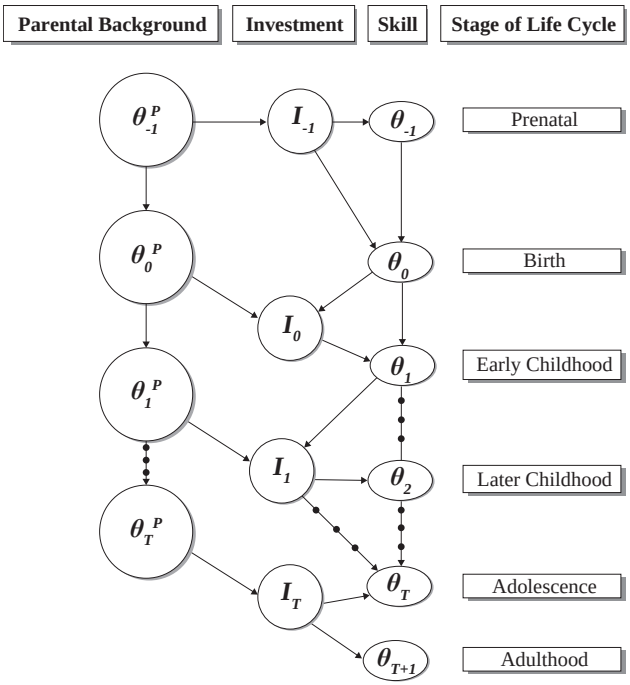
		Eligibility				Content				Sample Characteristics			Measures Available									
		Means-tested	High Disadvantage	Low Income	Criteria Narrowly Defined	Homogeneous Treatment	Medical Services	Home Visiting	Parent Involvement	Randomized Control Trial	Small Sample	Control Contamination	Age of Follow-ups	IQ	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.

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 1. At life cycle stage t, parental skills ($\boldsymbol{\theta}_t^P$), 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 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 (2014).

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 in period t complement each other and produce cognitive skills, non-cognitive skills,

[19] See Mayer (1997) and Heckman and Mosso (2014).
[20] Heckman and Kautz (2012, 2014).

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. (2010), 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]

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 (2014) summarize the literature.

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

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

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, 2014). 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 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

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

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.

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 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

[26] See Heckman and Vytlacil (2007).

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

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 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 Evidence from Demonstration Programs

This section analyzes the evidence from the demonstration programs listed in Table 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 3.2. We discuss common methodological issues that arise when analyzing these programs in Subsection 3.3. In Subsection 3.4 we present evidence on the short-term effects from these programs. We present evidence on long-term effects in Subsection 3.5. Subsection 3.6 relates the short-term findings to the long-term findings. Subsection 3.7 discusses cost-benefit analyses for two major demonstration programs, PPP and ABC. Subsection 3.8 summarizes the discussion.

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

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 use different eligibility criteria. The evidence on universal programs discussed in Section 5 shows that early childhood education is particularly effective for disadvantaged children.

[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.

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.

3.2 Overview of Programs Discussed in This Section

Table 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., 1982; Ramey et al., 1982; Zigler and Muenchow, 1994).[32]

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

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 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

[33] Appendix C provides further details about CARE.

[34] See Conti 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.

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 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	-	-	-	-
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 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: 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 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 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., 2010a; 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 (Elango 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. (2010a); 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.

3.4 Effects on IQ, Achievement Test Scores, and Conscientiousness

Table 4 presents estimated treatment effects on early IQ, early and late achievement test scores, and early conscientiousness pooled over genders. Tables 5 and 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. (2010a). 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 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. (2010a) 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 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 4.

Table 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 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 4 and 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, 2013). 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., 1982), Head Start (see Deming, 2009; Zhai et al., 2014), and state programs (see Lipsey et al., 2013).

Figure 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

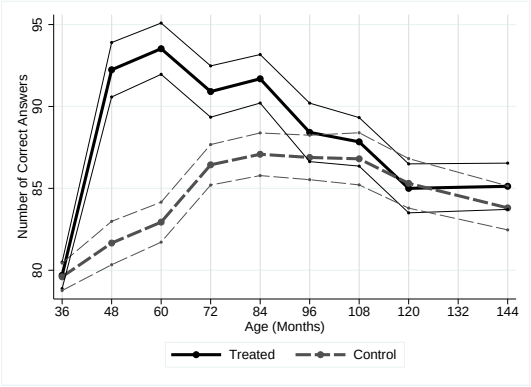
[39] See Heckman and Kautz (2012).

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 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.

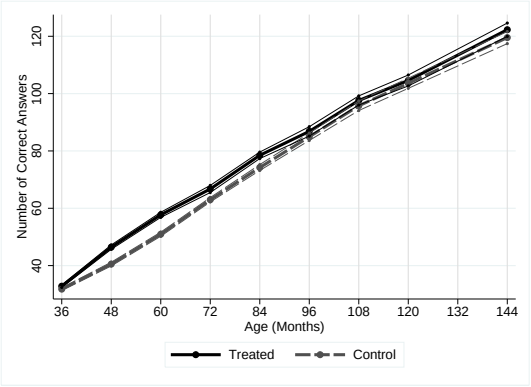
Hojman (2015) analyzes the causes of fadeout in cognition measured by IQ for PPP and ETP. He finds that the gains experienced by the treatment group occur rapidly during the first months of treatment and are followed by small or zero gains in the subsequent years of treatment. He also finds that almost 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 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 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 (2013) 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.

3.5 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 7, which is based on results from Heckman et al. (2010a, 2013); Campbell et al. (2014), and Elango 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 beneficial 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

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

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 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. (2010a). [b] Campbell et al. (2014). [c] Elango 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. (2010a) 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. $p - values$ 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 Elango 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.

3.6 Connecting Short-Term and Long-Term Effects

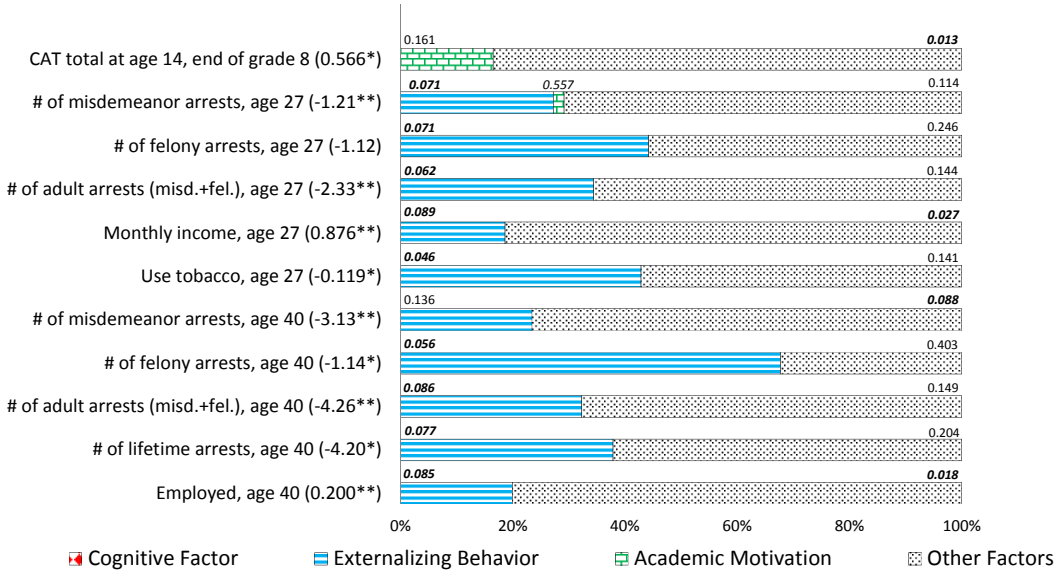
Dissipation of initial IQ gains is a common finding across programs. In some cases, IQ gains completely dissipate by the teenage years. Analysts focusing solely on IQ as a measure of program effectiveness confront a puzzle: Why do early childhood education programs have long-term effects if the effects on IQ dissipate? Heckman et al. (2013) 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. (2013) decompose the effects of PPP on later-life outcomes using a mediation analysis. The results of this are reported in Figures 3 and 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).

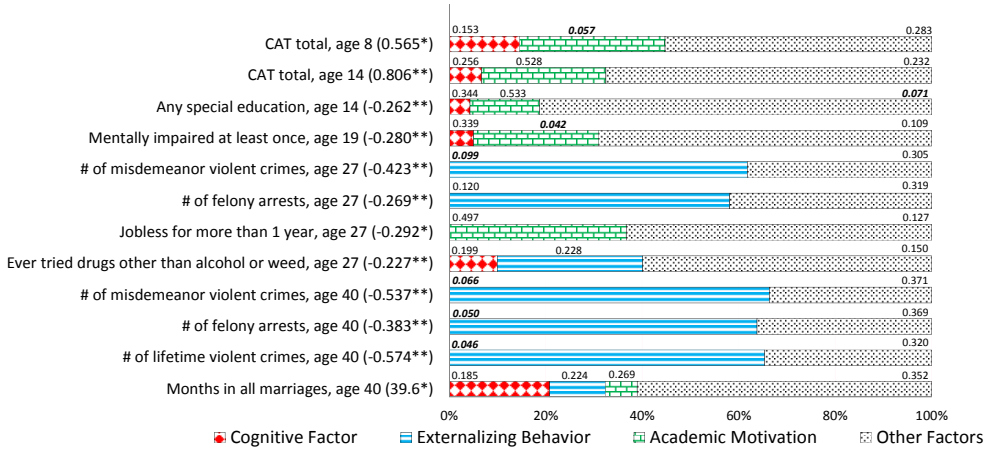
[43] See Heckman et al. (2013).

Figure 3: Decompositions of Treatment Effects of PPP on Male Adult Outcomes



Source: Reproduced from Heckman et al. (2013). 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. (2013) 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 4: Decompositions of Treatment Effects of PPP on Female Adult Outcomes

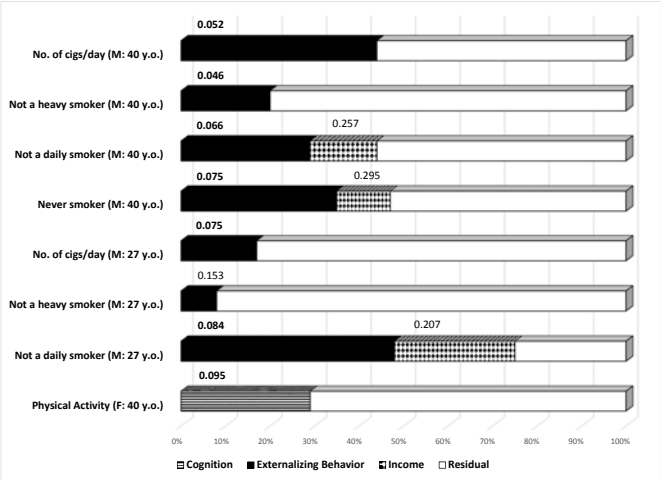


Source: Reproduced from Heckman et al. (2013). See note in Figure 3.

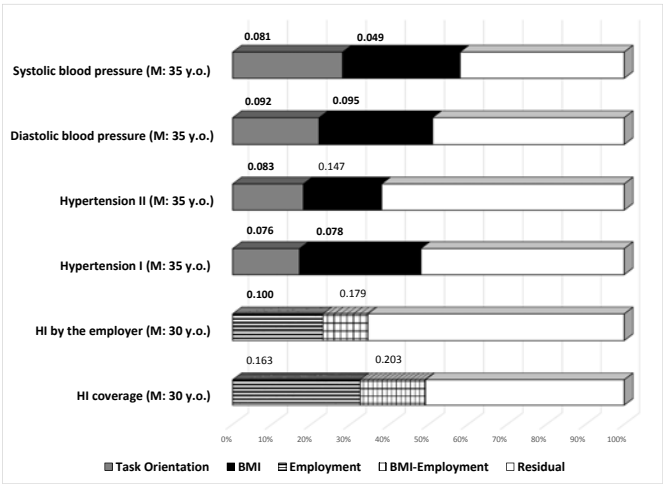
Conti et al. (2015) conduct a similar analysis for both PPP and ABC but focus on health outcomes. According to their findings, externalizing behavior is the primary mediator for the outcomes found in PPP, which is consistent with the findings in Heckman et al. (2013). For ABC, they find that task orientation and childhood BMI mediate approximately half of the improvements in blood pressure and hypertension found for males in the treatment group. Figures 5a and 5b illustrate the results from their mediation exercises.

Figure 5: Decompositions of Treatment Effects of PPP and ABC on Male Adult Outcomes

(a) PPP



(b) ABC

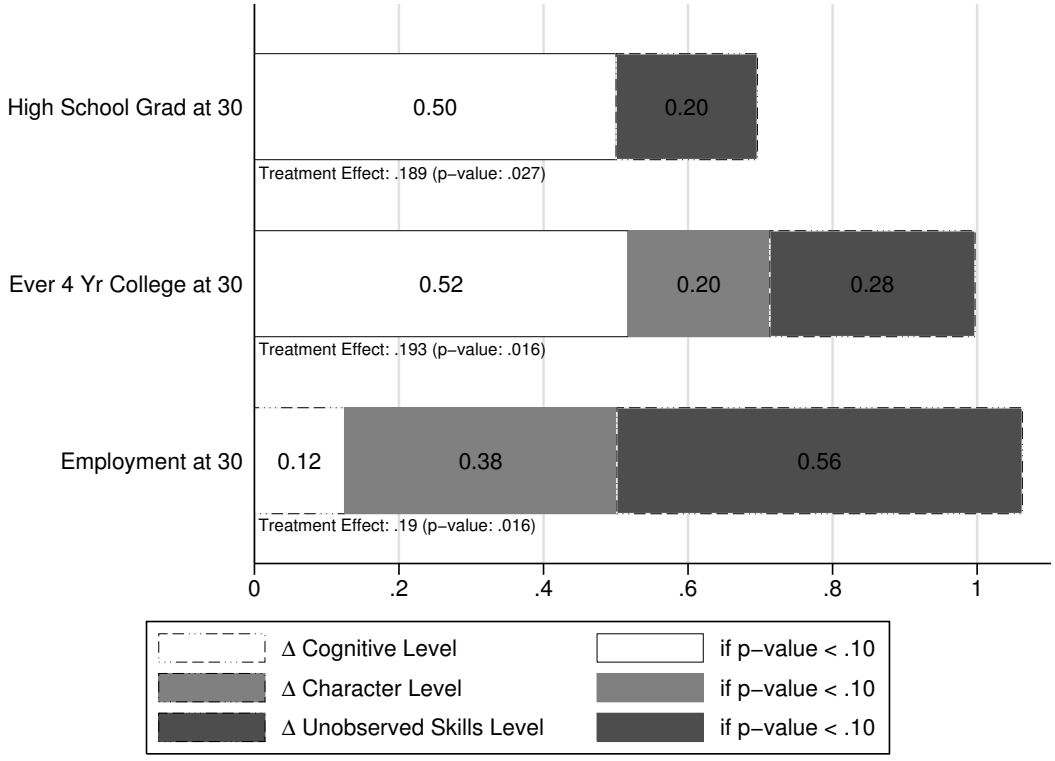


Source: Reproduced from Conti et al. (2015). Note: This graph provides a simplified representation of the results of the dynamic mediation analysis of the statistically significant outcomes for PPP and ABC. Each bar represents the total treatment effect normalized to 100%. One-sided $p - values$ that 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. (2013) among the early childhood inputs; and income as in Heckman et al. (2010a) 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 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 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.

3.7 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, 2001). 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. (2010b) 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. (2010b) 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. (2010b) 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 8.

Elango 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. (2010b).
[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 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 (Elango 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 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. (2010b); ABC estimates from Elango 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.

3.8 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 analysis suggest that provision of high-quality targeted programs can yield large returns on investment. These programs also provide childcare and facilitate working by the mothers of disadvantaged children.

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. Project Head Start (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. As a group, these studies are imprecise

[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).

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.

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

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

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 motor, language and literacy, cognitive, and social-emotional development. It emphasizes the importance of a supportive learning environment and the relationship between caretaker and

[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.

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 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.

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 demonstration 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 limita-

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

tions 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 9.

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 Administration for Children and Families (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

[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.

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. (2014) 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. (2014). 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. (2014); Kline and Walters (2014).

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

[59] We take these numbers from Feller et al. (2014). 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 instruments in this context.

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. (2014) (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. (2014) 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 effects for non-cognitive outcomes, but using parent reports, children generally become less

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

[62] Inverse probability weighting.

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.

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 (1995), 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 (1995) 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 (1995). 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. Garces et al. (2002) document positive effects on crime for African-American participants,

[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.)

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.

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 (1995) 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. Elango et al. (2015) present a more principled extrapolation analysis and a discussion of general procedures.

Table 9: Evidence Across Studies of the Impacts of Head Start

Study	Currie and Thomas (1995)	Garces et al. (2002)	Ludwig and Miller (2007)	Deming (2009)	Carneiro and Ginja (2014)	Feller et al. (2014)	Kline and Walters (2014)	Zhai et al. (2014)	Perry Preschool (Various sources)	Abecedarian (Various sources)
Dataset	C-NLSY	PSID	Multiple	C-NLSY	C-NLSY	HSIS	HSIS	HSIS		
Subpopulation	AA	AA, mother 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 (1995) originally report impacts on IQ/achievement in terms of test scores: PPVT at age 8 in Currie and Thomas (1995) 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 (1995) 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. (2014) 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. (2014), 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 Elango 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 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 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.

4.6 Summary of the Evidence from Head Start

We summarize the estimates for Head Start that are reported in the literature in Table 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]

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 centers, although it is not clear whether any of these are included in the program's evaluation.

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

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 treatment group was 4 percentage points less likely to repeat a school grade. Short-term effects on

[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.

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.

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, 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 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 of expanding the program to moderate-income children. Its goal is to increase the quality

[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.

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 5.1). Second, we summarize studies of a group of universal preschool programs in Oklahoma, Georgia, and Boston (Section 5.2). We then summarize the findings of the section (Section 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 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.

5.1 Universal Subsidies to Childcare

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. When they decompose results for a subsample of children of high school dropouts and high

[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).

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.

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. The policy caused a sizable increase in maternal labor supply (around 10 percentage points)

[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.

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.

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 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

[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.

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 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

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

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.

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.

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 (2006) 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.

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.

References

Administration for Children and Families (2005, May). Head Start Impact Study: First year findings. Technical report, U.S. Department of Health and Human Services, Washington, DC.

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.

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. 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.

Belfield, C. R., M. Nores, W. S. Barnett, and L. Schweinhart (2006). The High/Scope 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). Child care choices and children's cognitive achievement: The case of single mothers. *Journal of Labor Economics 29*(3), 459–512.

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 (2006). 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: Elsevier.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review 82*(2), 225–246.

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., 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.

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.

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

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

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.

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). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *Quarterly Journal of Economics 126*(4), 1593–1660.

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

Conti, G., J. J. Heckman, J. Yi, and J. Zhang (2015). Early health shocks, intrahousehold resource allocation, and child outcomes. Under Revision, *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 (2008, 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). 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 (2010, May). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

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

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

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 (2013). Investing in preschool programs. *Journal of Economic Perspectives 27*(2), 109–132.

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 (2013). 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., J. L. García, 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.

Feller, A., T. Grindal, L. Miratrix, and L. Page (2014). Compared to what? Varation in the impacts of early childhood education by alternative care-type setting. Harvard University, Department of Statistics.

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

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

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). The effects of universal pre-K on cognitive development. *Developmental Psychology 41*(6), 872–884.

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

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.

Havnes, T. and M. Mogstad (2011). 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. J. (1992). Randomization and social policy evaluation. In C. F. Manski and I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs*, Chapter 5, 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 impacts of early interventions. 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. and T. Kautz (2012, 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 Rothkopf, 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. Q. Yavitz (2010a, 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 (2010b, 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 (2014). 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 (2013). 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. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 71, pp. 4875–5143. 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.

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.

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). 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.

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.

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.

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.

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.

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.

Mayer, S. E. (1997). *What Money Can't Buy: Family Income and Children's Life Chances*. Cambridge, MA: Harvard University Press.

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., S. S. Aitken, and A. N. Smith (1985). *The Impact of Head Start on Children, Families and Communities*. Washington, DC: United States Head Start Bureau.

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.

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 (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.

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.

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.

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.

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

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., 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.

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.

United States Census Bureau (2014). American Community Survey. Technical report, United States Census Bureau.

U.S. Department of Education (2015). Preschool grants for children with disabilities: Funding status. Website.

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.

Weikart, D. P. (1967). Preliminary results from a longitudinal study of disadvantaged preschool children. 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.

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.

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 (2013). 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.