Early Educational Programs and Skill Development in the US*

Sneha Elango
The University of Chicago

Jorge Luis García
The University of Chicago

James J. Heckman
American Bar Foundation
The University of Chicago

Andrés Hojman
The University of Chicago

First Draft: May 30, 2014
This Draft: February 22, 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, Successful Pathways from School to Work, an initiative of the University of Chicago's Committee on Education funded by the Hymen Milgrom Supporting Organization, a grant from the Institute for New Economic Thinking (INET) supporting the Human Capital and Economic Opportunity Global Working Group (HCEO)—an initiative of the Center for the Economics of Human Development (CEHD) and Becker Friedman Institute for Research in Economics (BFI). We are very grateful to Joshua Ka Chun Shea for data managing, research assistance, and useful comments. We also thank Marianne Haramoto, Sylvi Kuperman, Ana Paula Melo, Jack Schmerold, Haonan Zhou, and Anna Ziff for research assistance. Carey Cheng, Fernanda Díaz de la Vega, Sindhu Gnanasambandan, and Damini Sharma helped us improve this draft with useful comments. We thank Robert Moffitt, David Blau and the other authors of this volume for valuable comments and insights. The views expressed in this paper are those of the authors and not necessarily those of the funders or persons named here or the official views of the National Institutes of Health.

Contents

1 Introduction

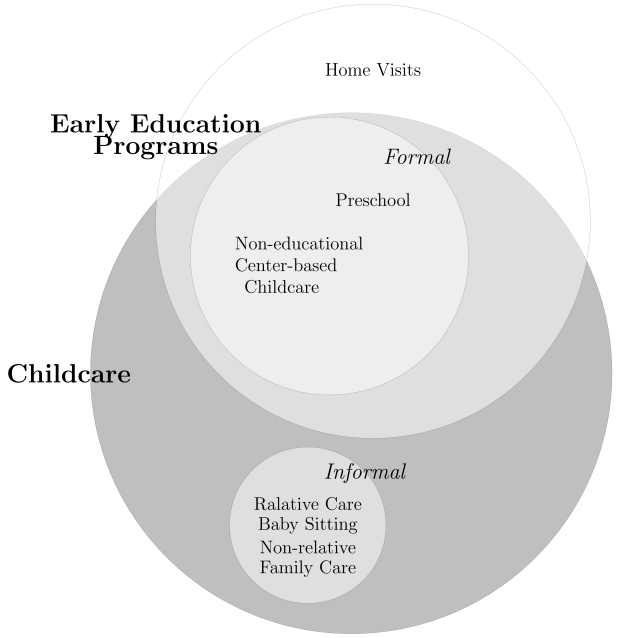
Table 1: Average Vocabulary and Number of Words Heard by Child per Hour, Age 3

Family Socio-Economic Status	**Actual Differences in Quantity of Words Heard**	**Actual Differences in Quantity of Words Heard**	**Child's Cumulative Vocabulary**
Welfare	616 words	5 affirmatives, 11 prohibitions	500 words
Working Class	1,251 words	12 affirmatives, 7 prohibitions	700 words
Professional	2,153 words	32 affirmatives, 5 prohibitions	1,100 words

Source: Hart and Risley (1995). Note: This table shows means for the speech and vocabulary development of children of various socioeconomic status by age 3.

Table 1 summarizes a classical study on maternal input and child skill output by socioeconomic status. Children born to families on welfare hear fewer words and have a less developed vocabulary. On average, children with parents performing professional activities have twice the vocabulary usage as children with parents on welfare. In this study, the differences appear as early as age 3. In addition, research on early childhood documents that these gaps persist into adulthood (see Heckman et al. (2013) and Heckman and Mosso (2014)). Thus, it is fundamental to study how early care environments affect child development to understand if there is a case for the public provision of Early Education Programs in the US. This is the aim of our paper.

Figure 1: Categories of Early Education Programs (EEPs) and Childcare



Note: this figure categorizes all of the types of programs we discuss. Formal refers to any program that is licensed and regulated by the state. Formal programs are typically center-based. EEPs that are childcare fall into this category; home visiting EEPs do not. Informal programs are unregulated by the state and include relative care, family care, and in-home care by a baby sitter or nanny. Relative care is care by a non-parent family member within or outside of the child's home. Family care is care provided by a non-relative family within or outside of the child's home.

We focus on Early Education Programs (EEPs) and study their impacts on skill development in early childhood. There is some overlap between EEPs and childcare arrangements; however, not all EEPs have childcare components (e.g. home visiting programs), and there are many childcare arrangements that lack educational components. We can further categorize childcare into formal and informal care arrangements. EEPs that function as childcare fall into the former category, which refers to any program that is licensed and regulated by the state. Formal programs are typically center-based. However, we cannot ignore informal childcare arrangements in this discussion, as they account for a substantial portion of the time children spend away from parents and are part of the counterfactual to EEPs. Informal programs are unregulated by the state and include relative care, family care, and in-home care by a baby sitter or nanny.[1] While we focus on EEPs in this paper, we conrtast their

[1]Relative care is care by a non-parent family member within or outside of the child's home. Family care is care provided by a non-relative family within or outside of the child's home. It is important to note that parents incur a monetary cost when using the latter and that this cost may be subsidized by federal or state grants.

impacts with those of other care options. See Figure 1 for a categorization of all of the types of programs we discuss.

Multiple surveys and reviews relate to our work. Several previous studies document the effects of EEPs (see Currie, 2001; Cunha et al., 2006; Blau and Currie, 2006; Duncan and Magnuson, 2013). Furthermore, Heckman (2008, 2012) and Cunha et al. (2010) formulate the case for early intervention based both on equity and efficiency arguments. Heckman and Mosso (2014) analyze the importance of early life conditions in shaping multiple life skills and state evidence on critical and sensitive investment periods for shaping different skills. They synthesize various structural models that analyze policies subsidizing the disadvantaged. Examples include subsidies to investment in early childhood (e.g. Cunha (2007)), restricted and unrestricted transfers (e.g. Del Boca et al. (2014a)), and increases in parental borrowing limits (e.g. Caucutt and Lochner (2011)). In general, the findings are positive and, therefore, establish a strong economic case for early intervention *for the disadvantaged* in order to improve child investment—suggesting theoretical and economic support for means testing (see Heckman, 2008). Lastly, Blau (2003) and Blau and Currie (2006) provide extensive surveys on childcare and its availability.

We see our work as a critical review of the current state of knowledge, bringing together facts and findings both on EEPs, childcare, and skill formation. Our conclusions indicate that EEPs have positive impacts on life-time outcomes through development of skills in early childhood, especially for disadvantaged children. We support this statement with 13 major findings from the literature that we elaborate on below and throughout the paper.

To do this, we first lay out a brief history of EEPs and major current policy questions in Section 2. Then, in Section 3 we describe the availability and take-up of childcare and document these main patterns: (F3-1) For the period 1990-2005, the take-up of formal childcare was around 50-60% and around 30% for informal childcare. (F3-2) There are clear gradients in the use of informal and formal childcare by family structure, parental education and labor force participation, and family income. (F3-3) Federal funding to childcare is proliferated

either to states that create regulations for expenditure or directly to organizations that oversee their own programs. (F3-4) More children are enrolled in state-funded preschool than in any other type of publicly funded program.

In Section 4 we focus on the main component of this paper: EEPs. We work with these programs because they have the richest data, due to the collection of yearly measures and availability of long-term follow-ups. We use evidence from this literature to establish the following empirical regularities: (F4-1) Demonstration programs strongly boost IQ in the short-run for disadvantaged children, and they may persist in the long-run. (F4-2) Demonstration programs boost non-cognitive skills in the short run (up through elementary school) for disadvantaged children. Their impact on later-life cognition is harder to measure. (F4-3) Demonstration programs are especially effective in boosting early and later life outcomes for disadvantaged children. (F4-4) Demonstration programs stimulate parenting and parent-child interactions. (F4-5) Cost-benefit analyses suggest that investing in Demonstration programs is economically efficient. We study randomized, controlled trials and observational studies of large-scale programs and note their design flaws and what we learn from them in Section 5. From them, we distill the following findings: (F5-1) Substitution bias is pervasive in recent program evaluations. (F5-2) The available evidence suggests that Head Start is not a failure, despite claims to the contrary. To finalize our empirical documentation we investigate the evidence on universal childcare and preschool in Section 6 and find that: (F6-1) Universal childcare policies have heterogeneous effects across children belonging to families from different socio-economic statuses. (F6-2) Impacts of programs are tied inextricably to program quality. In Section 7, we lay out some theoretical foundations of parental decisions and skill development. Then, we present evidence from structural models on these topics. We offer some final comments in Section 8.

2 Policy Questions and the History of Childcare and Early Education Programs in the US

In January 22nd of 2015, CBS news reported on President Obama's speech in the University of Kansas City. In short, he said, "we had universal childcare in the 1940's, so let's do it in 2015" (Condon, 2015). There are two major policy questions currently surrounding childcare and EEPs: (i) Are large-scale childcare programs beneficial for early childhood development? (ii) Should these programs be universal or targeted? We explore these questions throughout the paper. It should be noted that there has been a historical precedent for the public provision of childcare in the US, including EEPs and interventions. In this section, we briefly delineate this precedent and review the historical context in which EEPs emerged.

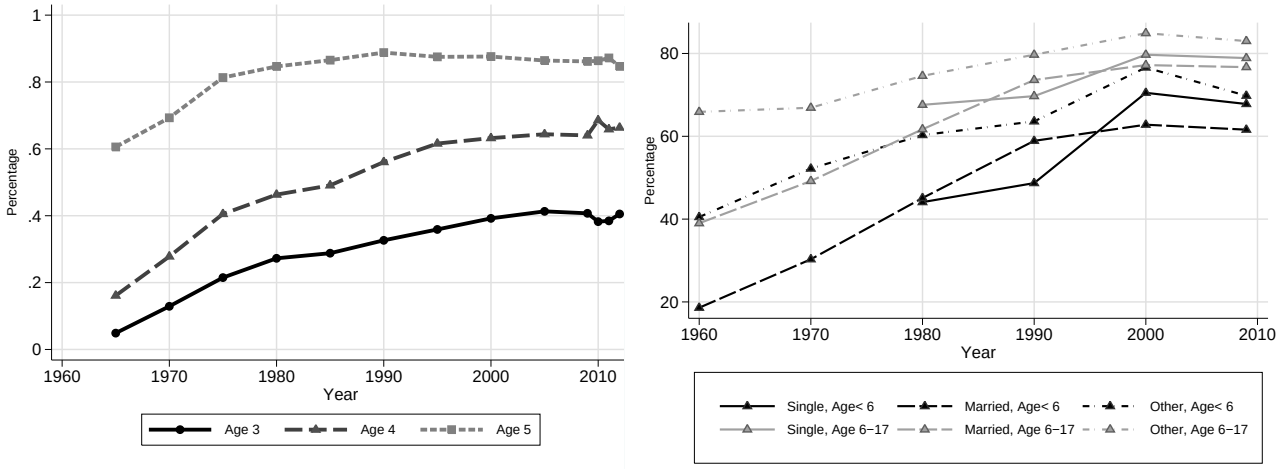
We identify on three main themes: (i) the availability of EEPs and childcare has historically been bolstered by the entry of women into the workforce (see Figure 2); (ii) many EEPs emerged as intervention-based programs in response to social and political changes; (iii) EEPs thus have a history of being targeted programs. In elaborating on these themes, we focus in particular on the historical context surrounding Head Start (the major large-scale EEP in US history) and the major demonstration programs we discuss in Section 4, the Perry Preschool Project (PPP), the Carolina Abecedarian Project (ABC), the Carolina Approach to Responsive Education (CARE), the Infant Health and Development Program (IHDP), and the Early Training Project (ETP).

Childcare was first popularized by the Industrial Revolution and the resulting urbanization and flow of women into the workforce. Nursery schools emerged in the early twentieth century through federal investment during the Great Depression and the influx of the women into the wartime work force (Shonkoff and Meisels, 1990). Not surprisingly, availability of childcare and EEPs has historically been tied to female labor force participation, and as Figure 2 this trend has continued for the last 50 years and can be expected to continue, especially as the number of single parent families increases (Ellwood and Jencks, 2004).

Figure 2: Maternal Labor Force Participation and Early Educational Program (EEP) Enrollment

(a) EEP Enrollment, Age 3-5

(b) Mother's Labor Force Participation Rates by Marital Status and Children's Age



Sources: EEP enrollment is created using the US Census Bureau; Current Population Survey, 2013. Maternal labor force participation uses Census 2012, Statistical Abstract. Note: In panel (b), other category includes widowed, divorced, or separated (including married, spouse absent).

Underlying this trend is a major disparity that has persisted through time. Historically, middle-income families have used preschool programs to supplement enrichment that their children receive at home. At the same time, populations that already experience additional constraints (e.g. low income families, particularly those headed by single mothers) have historically used childcare and similar resources to meet a need for childminding due to maternal employment (Cahan, 1989). Indeed, in a study on how to improve infant-toddler child care for low-income families, Paulsell et al. (2002) find that low-income mothers are facing multiple barriers in providing their children with high quality child care. These barriers can be separated into two categories: scarcity of high-quality child care for low-income families, and a lack of resources for low-income families to attend high-quality child care. In this context, EEPs, and especially interventions that target disadvantaged populations have come to fill an important need—subsidizing mothers' human capital, thus "remediating" the accident of birth, and reducing welfare dependency.

Addressing this need, many early EEPs were created as interventions. Figure 3 puts these programs into historical context. They typically served (i) low income populations

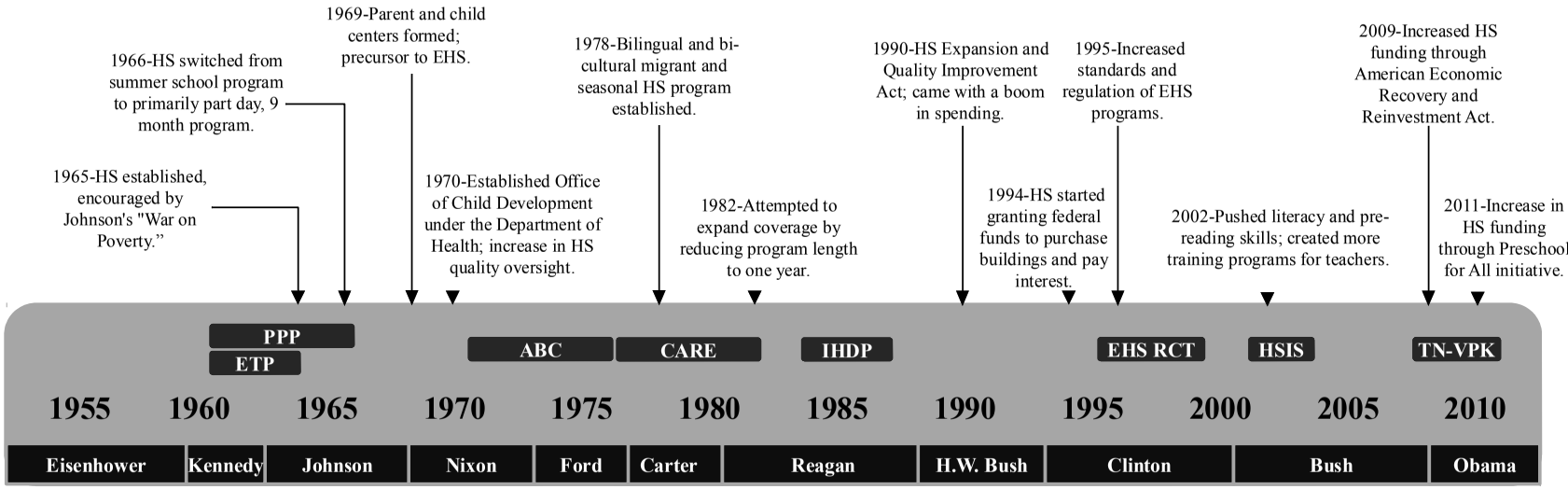
and (ii) groups marginalized by the social and political climate. Head Start was created shortly after the passage of the Civil Rights Act in 1964 as a part of President Lyndon B. Johnson's War on Poverty. ETP and PPP were also implemented on the heels of the Civil Rights Movement and targeted African Americans specifically. ABC and CARE occurred in the 70s and 80s, but served a primarily African American community. These programs occurred in small, even semi-rural towns, and can attribute some of their success to the progressiveness of program implementors (Chafe, 1981, p. 239).

Though they had target populations, HS, ETP, PPP, and ABC began with the relatively general goal of addressing risk for "mental retardation"[2] and promoting "school readiness". (Zigler and Muenchow, 1994; Gray et al., 1982; Weikart, 1967; Ramey et al., 1979). Over time, programs have become more targeted in their goals and more heterogeneous in the populations they serve. IHDP is the most recent of the demonstration programs we study, and it differs from the other programs in a few key ways. Firstly, it explicitly targeted particular at-risk populations (in this case, low birth weight infants). Secondly, it was implemented in relatively larger cities.[3]

[2]The clinical definition and understanding of mental retardation in the mid-twentieth century was much different than our understanding today, characterized by a conflation of "mental retardation" and risk factors associated with disadvantage that hindered early development.

[3]Little Rock, Arkansas; Bronx, New York; Boston, Massachusetts; Miami, Florida; Philadelphia, Pennsylvania; Dallas, Texas; Seattle, Washington; and New Haven, Connecticut (Brooks-Gunn et al., 1994a, p. 3035).

Figure 3: Head Start and Early Childhood Interventions in the US: History and Social Context



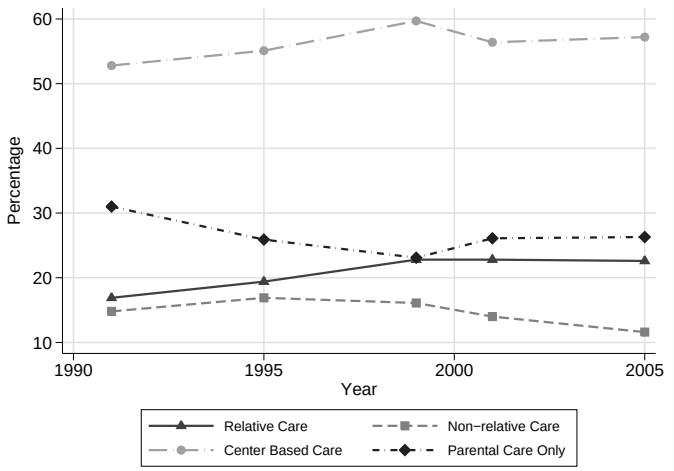
3 Childcare and Preschool in the US: Availability and Take-Up

In this section, we use various data sources to document the current availability and take-up of childcare and preschool in the US. We disaggregate the take-up of childcare into formal and informal care. It is important to note that while some EEPs are formal childcare, it is difficult to identify the quality of formal care arrangements as a whole, especially with respect to their focus on education. In fact, it is generally difficult to make broad comparisons of the take-up of EEPs because of the heterogeneity of quality and the difficulty of documenting and classifying their educational components. Thus, we are not able to describe the overall take-up of EEPs. Rather, we focus on the take-up of preschool and Head Start, which are well-documented and make up a substantial part of formal, center-based EEP use in the US. We explore the current state of (i) patterns in the use of informal and formal childcare arrangements; (ii) gradients in the use of childcare by different demographics; (iii) the availability of federal programs and federal funding streams; (iv) Head Start and preschool enrollment. We pay particular attention to Head Start because it has historically been the main driver of federal public policy directed to early intervention in the US.

F3-1 **For the period 1990-2005, the take-up of formal childcare was around 50-60% and around 30% for informal childcare.** We use data from the US Department of Education and limit ourselves to the classifications of childcare arrangements in their data to study take-up when children are not in kindergarten and are 3–5 years old. As defined in Section 1, formal childcare and center-based care are synonymous. However, some of these centers may not have a well-defined educational purpose, so this cannot be interpreted as the national take-up of EEPs. Figure 4 displays the percentage of children using each of the types of childcare. Importantly, in each year children might fall into more than one category, because they are taken care of in multiple arrangements. Center-based take-up is very steady around 55%. As defined

in Section 1, informal childcare is made up of relative and non-relative care. Their use seems fairly stable and accounts for a few percentage points more than 30% of the usage. Finally, around 22% of the children are taken care of exclusively by their parents. This is fundamental when analyzing the treatment effects of programs such as Head Start. Different families have different access and decision making processes with respect to outside options. It is not straightforward to interpret the impacts of programs in this situation—children in the treatment group are being compared both to individuals in the control group that are being taken care of by their parents, other comparable center-based arrangements, or by non-relatives in unregulated childcare arrangements. The child's potential outcomes in each of those situations could be substantively different. This is fundamental in the evaluation of childcare programs, as discussed in Section 6.

Figure 4: Primary Care Arrangements for Children Aged 3-5, not yet in Kindergarten



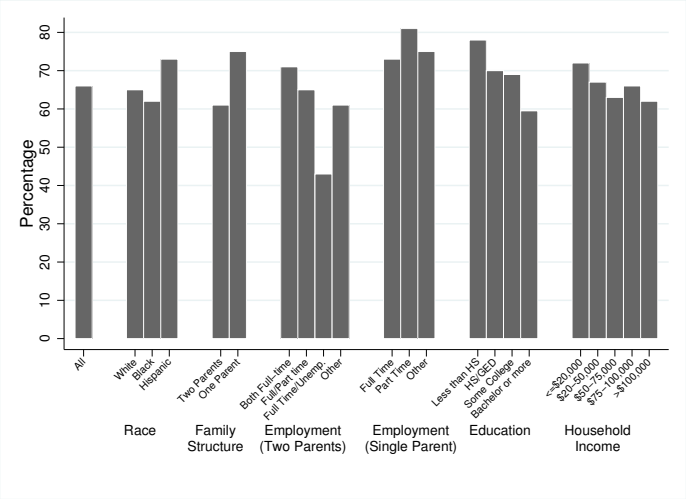
Sources: U.S. Department of Education, National Center for Education Statistics, Early Childhood Education Survey, Parent Survey, and Early Childhood Program Participation Survey of the National Household Education Surveys Program (ECE-NHES:1991; Parent-NHES:1999; and ECPP-NHES: 1995, 2001, and 2005); Note: a child's "primary arrangement" was defined as the regular non-parental care arrangement or early childhood education program in which the child spent the most time per week. Center-based programs include day care centers, nursery schools, prekindergartens, preschools, and Head Start programs. Non-relative Care includes Family Child Care (in another home) and Sitter (in child's home). Percents may not add up to 100%, because some children participated in more than one type of primary care arrangement.

F3-2 **There are clear gradients in the use of informal and formal childcare by family structure, parental education and labor force participation, and family income.** For both informal and formal care, there is no strong gradient by race, though

more Hispanics use informal care than formal care. Not surprisingly, more single parent households informal childcare than formal care, while approximately the same amount of two parent families use both types of care. More single parent families use informal care than two parent families, and this pattern is reversed for formal care. Furthermore, there is a very evident gradient in use by employment status for both single and two parent families. More two parent families with higher labor force participation use formal care and while fewer two parent families with higher labor force participation use informal care. The opposite pattern is apparent for single parent families. Furthermore, the evidence is favorable for relatively more advantaged households. As family income and parental education increase, more families use formal care. The opposite pattern appears for informal care. In both cases, the gradient on parental education is stronger than the gradient on family income. Additionally, there is an income gradient even within the take-up of different types of formal care. From Figure 7, we learn that low income families are much more likely to use Head Start than other preschool options.[4] For all groups, the use of formal care at large is about the same as the use of informal care. These levels differ from the use of care described in Figure 4 because Figure 4 show trends for children aged 3-5, while Figures 5 and 6 show enrollment for children aged 0-5.

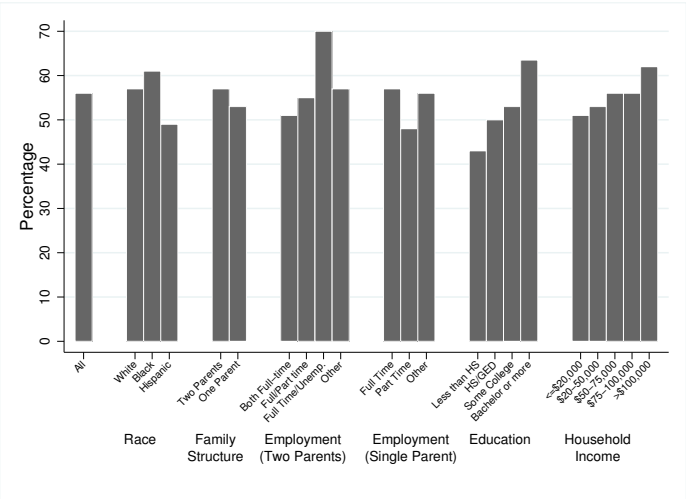
[4]High income families may not use Head Start because they do not meet eligibility criteria.

Figure 5: Informal Non-Kindergarten Childcare Enrollment for Children Aged 0-5, 2012



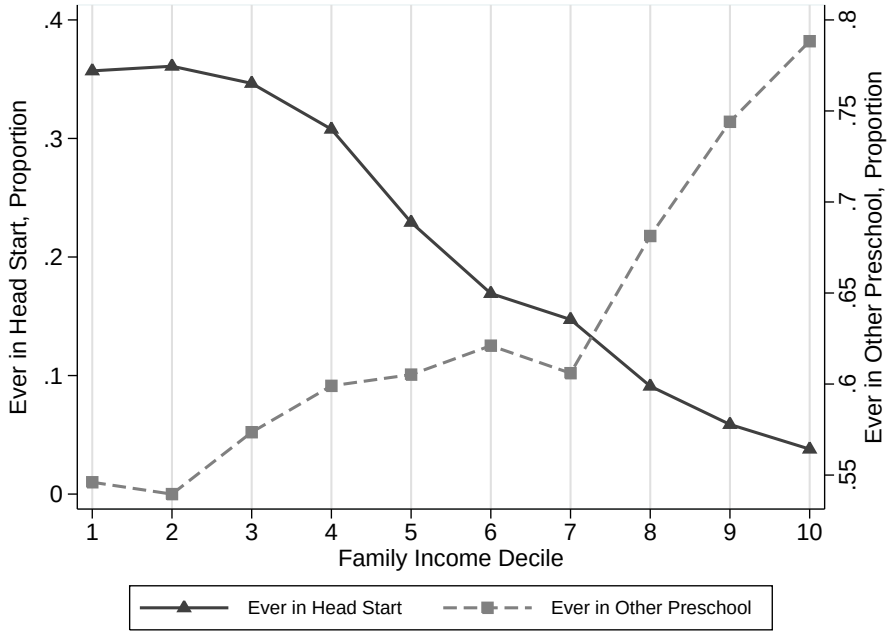
Source: U.S. Department of Education, National Center for Education Statistics, Early Childhood Program Participation Survey of the 2012 National Household Education Surveys Program (ECPP-NHES:2012); Note: this figure displays children from birth through age 5 and not yet in kindergarten participating informal care arrangements. It disaggregates the quantity of informal care used as primary care (i.e. used more than other forms of care) by families of children age 5 and under who are not yet in kindergarten. Informal child care includes relative and non-relative care (including babysitters). We do not account for children living separately from their parents when decomposing non-educational child care enrollment by family structure. We disaggregate by child's race, family structure, parent's employment status and highest education level and household income. Full-time employment is defined as working 35 or more hours a week; part-time employment is defined as work for less than 35 hours a week. For two parent families, "Other" refers to any other combinations of working status; for single parent families, "Other" is an average of parents looking for employment and parents not in the labor force. Education refers to the highest level of parents' education. HS stands for high school and includes GED. College is defined as having attended some college or having received an associate's degree or a vocational degree. The percentages are drawn over children with at least one weekly nonparental care arrangement. Children may have multiple arrangements.

Figure 6: Formal non-Kindergrten Childcare Care Enrollment for Children Aged 0-5, 2012



Source: U.S. Department of Education, National Center for Education Statistics, Early Childhood Program Participation Survey of the 2012 National Household Education Surveys Program (ECPP-NHES:2012); Note: this figure displays the percentage of children from birth through age 5 and not yet in kindergarten participating formal care arrangements. It disaggregates the quantity of formal care used as primary care (i.e. used more than other forms of care) by families of children age 6 and under who are not yet in kindergarten. Formal child care includes center-based care, including Head Start. We do not account for children living separately from their parents when decomposing non-educational child care enrollment by family structure. We disaggregate by child's race, family structure, parent's employment status and highest education level and household income. Full-time employment is defined as working 35 or more hours a week; part-time employment is defined as work for less than 35 hours a week. For two parent families, "Other" refers to any other combinations of working status; for single parent families, "Other" is an average of parents looking for employment and parents not in the labor force. Education refers to the highest level of parents' education. HS stands for high school and includes GED. College is defined as having attended some college or having received an associate's degree or a vocational degree. The percentages are drawn over children with at least one weekly nonparental care arrangement. Children may have multiple arrangements.

Figure 7: Enrollment in Head Start and Other Preschool Options



Source: cNLSY. Note: this graph shows Head Start enrollment compared to enrollment in other preschool options. This is an updated version of the graph in Garces et al. (2002). The cNLSY tracks children of mothers in the NLSY79. The NLSY79 contains individuals who were age 14 to 22 in 1979, and it is nationally representative of that cohort. We show income in the household of the child in 2000.

F3-3 **Federal funding to childcare is distributed in two ways: (i) states receive federal funding and have freedom to allocate it under regulations of varying levels of restrictiveness or, (ii) federal grants are awarded directly to organizations.** Head Start (HS) and Early Head Start (EHS) are the most widely known federal programs that provide early childhood services, and they share part of their eligible populations with other federal programs. The major components of federal funding are the Preschool for All initiative, the Child Care Development Fund (CCDF), also known as the Child Care and Development Block Grant (CCDBG), and Individuals with Disabilities Education Act (IDEA) Preschool Grants.[5] Table 2 describes these funding streams. Of these programs, HS and EHS are unique. Though they have a stronger curricular component and a more holistic approach in serving

[5]Race to the Top is an additional source of funding for some states. States compete on the basis of the quality, outcomes, and progress of their programs; they are selected for awards between 37.5 and 75 million USD 2014.

whole families rather than individual children, the nature and quality of individual programs are so heterogeneous that we prefer not to think of Head Start as being a single program.

Table 2: 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 referral), 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 need to pass teacher/staff qualification req. and program quality and compliance evaluations.	2013 Federal Appropriation (including local projects and support activities): \$7,573,095,000. (2013 USD). 2013 Funded Enrollment (including Migrant programs): 903,679.
Early Head Start, 1994-present	Expectant mothers up to children aged 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 referral), 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 need to pass teacher/staff qualification req. 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 allocated to states that subsidize families to subsidize 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: \$5,002,940,470. (2013 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 provided to states based on proportion of disabled children 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: \$353,238,000. (2014 USD.) 2014 enrollment: 749,971 children.

Sources: **HS and EHS** : Vogel et al. (2006), Love et al. (2002), Administration for Children and Families (2009), and Administration for Children and Families (2008). There are some exceptions (up to 10% of enrollees) to the income requirements for special needs children and families that belong to tribes. **IDEA**: U.S. Department of Education (2009) and U.S. Department of Education (2012). **CCDF**: Administration for Children and Families (2013). 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 the 70s.

We exclude Preschool for All from Table 2 because we think of it not as a concrete funding strategy but as a comprehensive and broad initiative. It provides $75 billion dollars over 10 years targeting low-income[6] 4-year-olds, with the aim of expanding to moderate-income children (The White House, 2013). The goal is to increase quality and quantity of available preschool and to support voluntary home visiting programs for the most disadvantaged families by providing grants to states to expand their existing preschool infrastructure and Head Start options. Quality standards are to be measured by teacher qualifications, class size, health services, and assessment. The policy has a broad reach to grow the supply of existing programs, incentivizing full-day kindergarten programs, explicitly supporting HS and EHS, and creating a new EHS-childcare partnership to meet the childcare needs of working families.

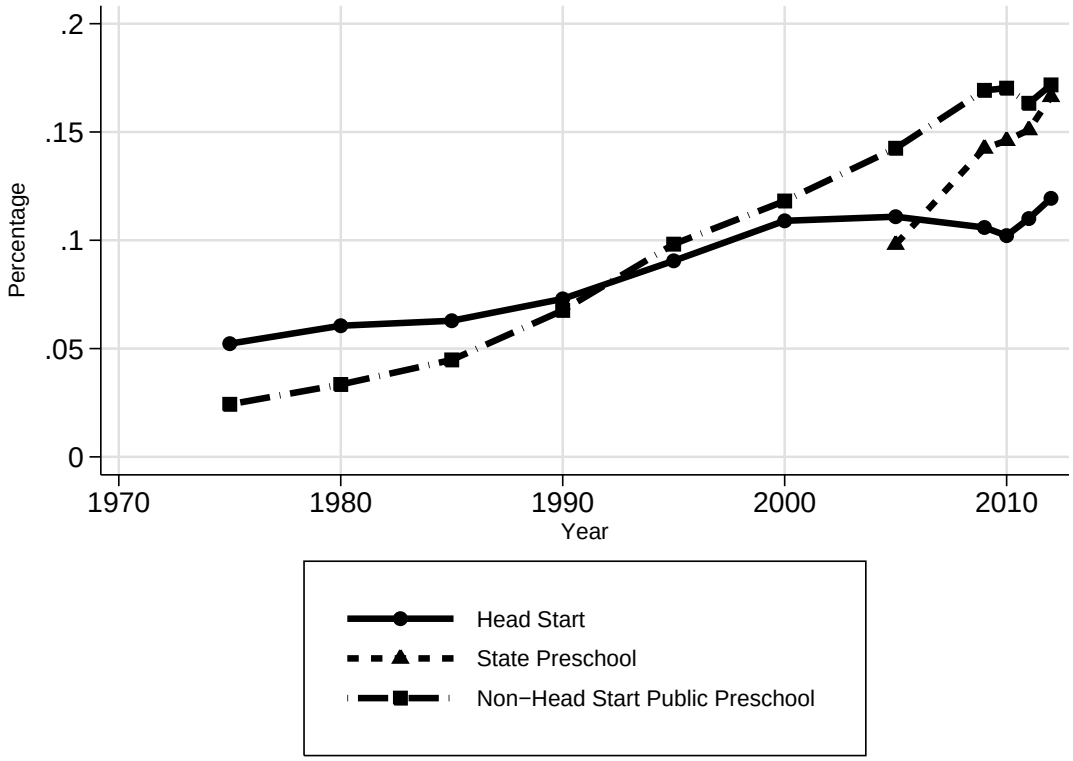
F3-4 **More children are enrolled in state-funded preschool than in any other publicly funded program.** States have their own programs with substantial funding coming from state resources. In recent years, this funding has grown most for financing state preschool programs. National Institute for Early Education Research (NIEER) studies show that more children are enrolled in state-funded preschool than in any other publicly-funded childcare arrangement: 28% of 4-year-olds are enrolled in state-funded programs, 11% in HS, 3% in other public programs, and 3% in special education (NIEER, 2013; Barnett and Yarosz, 2007).[7] Figure 8 contrasts the use of state and federal-funded preschool arrangements for 3- and 4-year olds. There is a strong increasing trend in the take-up of Head Start and public preschool. Beginning in the early 90s, public preschool began to surpass Head Start in take-up and has grown to surpass it by almost 50%. Non-Head Start public preschool includes state preschool, and it is important to note that the trends were constructed from three different data sources. Thus, the gap between state and non-Head Start public preschool may be

[6] Low-income is defined as being at or below the 200 % federal poverty level.

[7] This does not include special education children who are also enrolled in state-funded preschool or Head Start.

attributed to incongruities in the data. However, it is reassuring that there is no gap in 2012 between the two.

Figure 8: Percentage Three and Four Year Olds Enrolled in State Preschool



Head Start Source: Administration for Children and Families, Office of Head Start Annual Report 2011. Non-Head Start Public PreK Source: Census Bureau, Current Population Survey (CPS), October, 2012; Administration for Children and Families, Office of Head Start Annual Report 2011. State PreK Source: National Institute for Early Education Research State Preschool Yearbooks 2003-2013. Note: this figure shows the national pooled enrollment of 3- and 4-year-olds as a percentage of total 3- and 4-year-old population. Non-Head Start Public PreK is constructed by subtracting Head Start enrollment from CPS estimates of national public preschool enrollment. Non-Head Start Public Preschool includes state preschool.

4 Randomized Controlled Trials of Demonstration Programs

Several previous studies in Economics document the effects of randomized controlled trials of demonstration programs (see Currie, 2001; Cunha et al., 2006; Blau and Currie, 2006; Duncan and Magnuson, 2013). The most well-known of these programs are the Perry Preschool Project (PPP) and the Carolina Abecedarian Project (ABC). In this section we also discuss the Carolina Approach to Responsive Education (CARE), the Infant Health and Develop-

ment Program (IHDP), and the Early Training Project (ETP). We describe these programs in detail and review their evaluations. These programs are the highest quality of any of the programs that we cover in this paper. Four of the programs were center-based (with all but ABC having a home-visiting component) and one program (ETP) was a home-visiting program. Duncan and Magnuson (2013) present a meta-analysis of the impacts on IQ (or achievement) for 84 interventions. They find that (i) the impacts of PPP and ABC are on the high end of the spectrum of effect sizes (average .35 standard deviations near the end of the program); (ii) programs designed by researchers had higher effect sizes; (iii) older programs had higher effect sizes, which they attribute to higher-quality home environments and increases in other forms of center-based care; and (iv) there is no clear pattern in the duration of program or starting age. They explain that the evidence on program components is still weak, but that there are hints that teaching quality might be more important than structural variables (class size, teacher education, etc.). We generally share the perspective of these previous studies but substantially expand their analyses. In this survey, we dig deeper and explicitly discuss the differences in the curricula and inputs of these demonstration programs and cite new evidence about their importance. Importantly, we describe characteristics of their implementation that pose unique econometric challenges in their evaluations.

First, we describe each program and its components. Then, we survey the evidence from the literature. After reviewing all of the programs, we propose a methodology to perform our own evaluation and provide evidence supporting the findings in Section 4.9.

Only a handful of studies may be classified as long-term RCTs. To the best of our knowledge, the only two experimentally evaluated demonstration programs implemented in the US that have been followed for enough time to present reliable estimates of their labor market effects are ABC (and the closely related CARE project) and PPP. Through their long-term follow-ups, we can measure their impacts on labor market outcomes, crime, and health in adulthood. A few other demonstration programs have followed individuals after they finish school, but many of their participants were still enrolled in educational

programs when last surveyed. Although we cannot learn about these programs' impacts on labor market outcomes, they allow us to study an array of relevant results on cognitive and non-cognitive development. To the best of our knowledge, the only two experimentally evaluated educational demonstration programs implemented in the US with adequate follow up to consider these outcomes are IHDP and ETP. In this section, we present results using comparable methodologies and samples, and correct for several statistical problems that are present in the earlier literature (elaborated in Appendices B and C). As we understand it, all other US-based RCTs of early educational demonstration programs with long-term follow-ups have design or attrition problems that make them unsuitable for obtaining conclusions about long-term impacts.

We begin by summarizing and comparing PPP, ABC, CARE, IHDP, and ETP in Table 3. We also summarize the baseline characteristics of the sample for each demonstration program in Table 4.

Table 3: High-quality Early Childhood Interventions, Summary Table

	PPP	ABC	CARE	IHDP	ETP
Program Overview[1]					
Implemented years	1962–1967	1972–1982	1978–1985	1985–1988	1962-1968
Site	Ypsilanti, Michigan	Chapel Hill, North Carolina (UNC)	Chapel Hill, North Carolina (UNC)	8 sites selceted after competitive review	Segregated black schools in Abbotsfield, Tennessee
# Cohorts	5	4	2	1	2
N (Treatment:Control)	123 (58 : 65)	111 (57 : 54)	64 (41 : 23)	985 (377 :608)	88 (43 : 45)
Age of Entry	3–4	0	0	0	3–4
Duration	1–2 years	5 years	5 years	3 years	2–3 years
Treatment					
Home visits[2] (per month)	4	0	2.5–2.7	1.5	4
Center care (weeks per year)	30	50	50	50	10
Center care (hours per week)	12–15	45	30	20+	20
Parent Involvement		✓	✓	✓	
Nutrition		✓	✓		
Diapers/Child Care Goods		✓	✓	✓	
Health Check-ups		✓		✓	
Medical Care		✓	✓	✓	
Counseling			✓	✓	
Parenting Instruction	✓		✓	✓	✓
Control[3]					
Nutrition		✓(Formula up to 15 mo)	✓		
Diapers		✓(up to 115 mo)	✓		
Health Check-ups		✓(Cohort 1, up to Age 1)		✓	
Medical Care				✓	
Randomization Protocol[4]					
Steps	1. Rank by initial IQ 2. Group evens and odds 3. Balance gender, SES, etc. 4. Randomize whole group	1. Match on HRI 2. Adjust by gender, maternal IQ, siblings 3. Randomize	Same as ABC	1. Stratify on birthweight and site 2. Randomize	Simple randomization into 2 treatment and 1 control groups
Compromises	Siblings receive same assignment Working moms switched to control	2 extremely needy switched to treatment 4 refused random assignment			
Counterfactual	Stay at home	Stay at home or childcare	Treatment substitution problem	Stay at home or childcare Treatment substitution problem	Stay at home
Program Eligibility[5]	Cultural Deprivation Scale < 11 Low IQ (< 85) African American	High Risk Index > 11 Biologically healthy No signs of mental retardation	High Risk Index > 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[6]					
Adult-Child Ratio	1:5–1:6	1:4–1:6	1:3-1:6	1:3–1:4	1:4–1:6
Staff & Certifications					
Teachers	B.A.$^{\oplus}$	HS grad–M.A.; experience with kids	HS grads$^{\circ}$	College grads	$^{\circ}$
Specialists	Special Ed. Teachers$^{\oplus}$	Physician, Nurse Social Workers	College grads$^{\circ}$ M.A.$^{\circ}$	Education directors, M.A. Clinical staff	Teaching Assistants, college & PhD students Home visitors$^{\oplus\circ}$
Costs (2014 USD)	$20,911	$88,737		$22,187	

Sources: **PPP**: (Weikart et al., 1964, 1978; Weikart, 1970; Schweinhart et al., 2005). **ABC**: (Campbell and Ramey, 1994; Ramey et al., 1979; Masse and Barnett, 2002). **CARE**: (Wasik et al., 1990; Burchinal et al., 1997; Campbell et al., 2008). **IHDP**: (Brooks-Gunn et al., 1994b,a; McCarton et al., 1997; Ramey et al., 1992). **ETP**: (Gray and Klaus, 1965; Klaus and Gray, 1968; Gray and Klaus, 1970). Notes: **[1]** In ABC, at Preschool period: Group E consists of 57 children who received a 5-year preschool intervention. There was another randomization for a school-age program that is outside the scope of this paper. At Preschool period: Group C consists of 54 children who received no preschool intervention. In the treatment group of CARE, 15 children received both child care services and family education (referred to as Ch+H); the 26 remaining children received only family education (referred to as H).The two cohorts of CARE lived in or near a Southeastern university town, and were born between the spring of 1978 and early 1980. CARE originally selected 65 families with 67 children. 64 children are included in the final sample. 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. In ETP's treatment group, 22 of the children received three rounds of summer preschool, in addition to three years of weekly meetings with home visitor (these children are referred to as T1) ; the remaining 21 children in the treated group received two rounds of summer preschool, in addition to two years of weekly meetings with a home visitor (these children are referred to as T2). T1 received their first year of summer school at age 3. T2 received their first year of summer preschool at age 4. **[2]** In PPP, home visits were intended to involve the mother in the educating the child, increase her understanding of the educative 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. In CARE, the reported figures are the amount of home visits: 2.5 visits per month for the H; 2.7 visits per month for Group Ch+H. The original design was to have weekly home visits. 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. In ETP, T1 parents received two 9-month training sessions, while T2 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). **[3]** The control group of the first cohort ABC received health check ups for the first year, after which this practice was discontinued. **[4]** The randomization protocol was the following. PPP: (1) Matched on C.D. scale and SB scores; (2) Rank by IQ, separate even and odds; (3) Randomly assign even and odd to treatment and control. ABC: (1) Identified candidates before birth; (2) Pair matched on gender, maternal IQ, number of siblings, high risk index (the High Risk Index used in ABC and CARE is calculated through a weighted average of parents' age, education levels, family income, mother's IQ, father absence, poor school performance of siblings and seven other factors); (3) Randomly assigned one to control, other to treatment. Care randomly assigns 65 families using the ABC protocol. IHDP: Randomized by the National Study Office using an adaptive randomization model. IHDP balanced the treatment groups across two birth weight strata: higher low birth weight (2000g–2500g), and lower low birth weight (< 2000g). ETP: 61 Children from Abbotfield, TN were randomized into three groups, two treatment groups and one control group. **[5]** In PPP, criteria for home environment included education of parents, occupational level of father, maternal employment, and household density. In ABC and CARE 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 tests'; 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. **[6]** $^{\circ}$ signifies that staff were specially trained for the program. $^{\oplus}$ signifies that staff were state certified.

Table 4: Background Characteristics at Baseline, All the Programs

	PPP		ABC		CARE		IHDP		ETP	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Black	1.00	0.00	0.97	0.16	0.85	0.36	0.37	0.48	1.00	0.00
IQ	79.02	6.44	85.10	9.19	89.41	12.15	88.00	20.16	84.74	20.24
Mother's Age	29.10	6.57	19.89	4.82	21.68	4.73	24.87	6.00	29.71	8.49
Mother's Education	9.42	2.20	10.23	1.84	10.97	1.54	12.41	2.43	9.02	2.56
Mother Works	0.20	0.40	0.73	0.45	0.31	0.47	0.34	0.47	0.44	0.50
Father at Home	0.53	0.50	0.29	0.46	0.23	0.42	0.56	0.50	0.86	0.35
Father's Age	32.81	6.88	23.21	5.91	24.46	5.95	27.64	6.67	32.27	9.06
Father's Education	8.60	2.40	10.95	1.76	11.03	1.76	13.16	2.89	9.45	2.80
Father Works	0.86	0.35	0.87	0.34	0.82	0.39	0.57	0.50	0.98	0.15
# Siblings	4.28	2.59	0.64	1.10	0.66	0.94	1.02	1.17	3.51	2.16
Treatment	0.47	0.50	0.52	0.50	0.42	0.50	0.39	0.49	0.39	0.49

Source: Own calculations. Note : initial sample sizes are: **PPP**: 123; **ABC**: 111; **CARE**: 64; **IHDP**: 985; **ETP**: 88. 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. We generate an indicator variable that evaluates to 1 if the mother is working at the time of program entry, and 0 otherwise. We do the same with the fathers. We also generate an indicator variable that evaluates to 1 if the father lives at home at the time of entry, and 0 otherwise. 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. **CARE**. 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 prior to entry into summer school, and we measure this using the Stanford Binet Intelligence Scale. Mother's age is reported at the time of program entry, when the child is age 4. Test scores are constructed to have a national mean of 100 and a standard deviation of 15.

4.1 Perry Preschool Project (PPP)

4.1.1 Program Description

PPP might be the most well-known experimental preschool program; it has been followed for 40 years now and its rates of attrition are exceptionally low for this type of program.[8] Table 3 shows that the program focused on disadvantaged black children aged 3-5, and only ran for two years for around 3 hours every day. Although not all teachers had college degrees, they had experience and expertise in teaching. The student-teacher ratios were around 6:1, which is not unusual for high-quality preschool education programs. It also included weekly home visits by staff members. An important characteristic of PPP is that children in the control group had no access to substitute preschool programs. This could be part of the explanation to the relatively higher impacts that are observed in PPP compared with other programs in this survey (Weikart, 1970).

[8]Small departures of the randomization were assessed in Heckman et al. (2010a); the attrition rates were less than 9% at age 40, which is much less than usual for demonstration programs of this type.

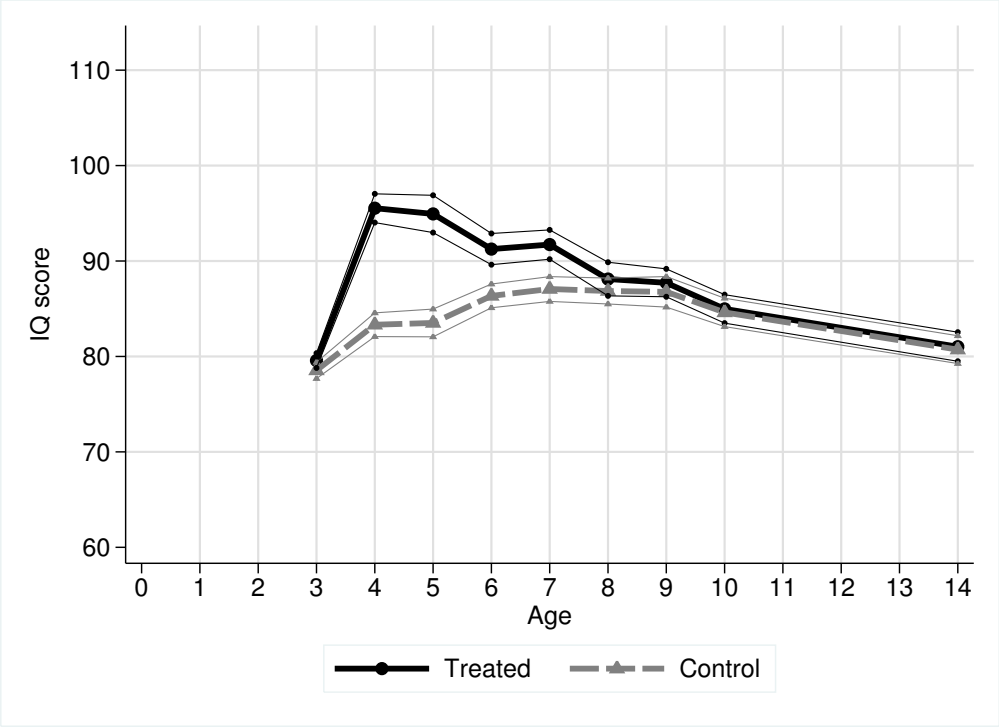
4.1.2 Evidence from the Literature

PPP has been evaluated by many analysts. Here, we rely on Heckman et al. (2010a) for adult outcomes. That paper (i) accounts for compromises in the randomization of the program, such as non-random switching between groups; (ii) uses inference methods that control for the number of variables tested, thus being robust to multiple hypotheses testing; (iii) uses permutation tests that are tailored to the randomization procedure of the program and that are exact, even in small samples.

Heckman et al. (2010a) report results by gender. For females, they find impacts of 56% on high school graduation and an impact on years of education completed of a little more than one extra year. They also find that the program reduces the total number of arrests up to age 40 by almost three. The probability of being unemployed during the last year at age 27 is reduced by 30%. No significant impacts on health or earnings are found. For males, they find no effects on education or self-reported health, but in terms of crime, they find a large reduction of almost 5 lifetime arrests. Treatment group individuals are 17% less likely to have been unemployed during the last year at age 30. Finally, monthly earnings in the current job increase by a little more than $1,000 in 2006 dollars at age 27 for the treated group, which is an increase of 70% compared to the control group.

Weikart (1970) gives an early analysis of the impacts on cognitive and non-cognitive skills. In terms of non-cognitive skills, that paper presents a few positive results, but lacks a coherent measurement scheme. It also describes for the first time the well-known pattern of impacts on IQ of PPP: there were very substantial impacts of the program in all types of IQ tests during the time children participated in the preschool program. In the case of PPP, the magnitude of those impacts was around a whole standard deviation. However, during the first years of elementary school, those impacts had dissipated to the point where there was no significant difference between the groups. This is one of the clearest known examples of the fadeout phenomenon, which is widely observed in preschool demonstration programs. This is illustrated in Figure 9.

Figure 9: Cognitive Skills Dynamics in the Perry Preschool Program



Source: 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. Cognitive skills have been measured by the Stanford-Binet IQ test in each year, which is standardized, based on a national sample, to have a US national mean of 100 points and a US national standard deviation of 15 points. PPP had a total of 123 participants: 58 treated and 65 controls.

These puzzling patterns sparked a new question in the literature: why did this program have so many substantial impacts on adult outcomes if the program did not have long-lasting impacts on IQ? This question was recently tackled by Heckman et al. (2013). In that paper, the impacts of PPP are decomposed to study the different contributions of different gains in skills to the final impacts on outcomes, using mediation analysis. They find that a substantial part of the impacts of the programs are explained by previous gains on non-cognitive skills. In particular, they find that for males, 40% of the total reduction in lifetime arrests and 20% of the reduction of unemployment at age 40 are explained by gains in externalizing behavior. On the other hand, for females, 30% of the gains in achievement tests, and 40% of the decrease in unemployment at age 27 are explained by gains in academic motivation. Consistent with males, up to 65% of the decrease in lifetime violent crimes is explained by a reduction in externalizing behavior.

The costs and benefits of PPP are evaluated in Heckman et al. (2010b). This report improved on several aspects of previous CBAs, including accounting for compromised randomization, presenting standard errors for the rate of return estimates, and accounting for deadweight loss. For comparability with other programs, we use no deadweight loss when reporting their results. We present disaggregated results for this paper in Table 6. The results are very positive, with a benefit-cost ratio of 8.6.[9] Benefits to the parents of the children (free child care, more labor market participation) and benefits to health are not considered, and they form a substantial part of the benefits in the case of other programs, so this evaluation might be conservative in estimating benefits.

4.2 The Carolina Abecedarian Project (ABC)

4.2.1 Program Description

ABC is a well-known and highly studied demonstration program that took place between 1972 and 1982 in Chapel Hill, NC. ABC is one of the most intensive of the experimental demonstration programs we study. It provided day-long high-quality childcare and medical services from ages 0-5. Additionally, children were transported to the program site by bus drivers and received two meals and a snack daily, along with regular health check-ups (Ramey et al., 1979; Campbell and Ramey, 1994).

4.2.2 Evidence from the Literature

The Abecedarian Project data has been used in dozens of papers. It is the only major randomized demonstration program that had general long-lasting impacts on IQ (Hojman, 2015). For example, Campbell et al. (2002) report a significant difference between treatment and control groups of 5 IQ points at age 21. The last general evaluation of adult outcomes is Campbell et al. (2012). They find positive impacts of 1.2 extra years of education; an

[9] A very highly cited finding of this paper, the 7:1 benefit-cost ratio, is calculated using positive deadweight loss estimate.

increase of 17% in high school graduation; a non-significant difference in yearly earnings of $16,803 2014 dollars per year (50% more than the control group mean); and no impacts on crime.

Particularly useful about ABC data is that around age 34, a new round of data was gathered and medical tests were performed on program subjects. The results are discussed in Campbell et al. (2014a) and Campbell et al. (2014b). The sample for this study is smaller than in the normal ABC data due to an attrition rate of around 30%. Using testing procedures that are exact in small samples; controlling for size in multiple hypotheses groups; and correcting for attrition, they find significant treatment effects for males in blood pressure, hypertension and various measures of combined heart risk. For females, they find impacts on abdominal obesity (21% difference), one combined risk measure and on one measure of prehypertension.

An interesting characteristic of ABC (and CARE, see next section) is that among the demonstration programs that we study, it had the highest rate of treatment substitution. At age 5, the average number of months in preschool for the control group was 7.5 in a year. The preschools seem to have been of relatively high quality (Burchinal et al., 1989). While the substitution rate was much smaller for ages 0-2, it is likely that this treatment substitution led to smaller estimated impacts for the program are than what would have been obtained if no control child had gone to a preschool, as in the cases of Perry and ETP.

Barnett and Masse (2007) evaluate the costs and benefits of the ABC program. Benefits are estimated from (i) earnings from participants; (ii) earnings from future generations; (iii) maternal earnings; (iv) school education savings; (v) improved health; (vi) higher education costs; and (vii) welfare use. All of these benefits are compared with those of PPP in Table 6. The costs of the program are calculated using a mix of reports of the implementation of the program and estimates for the costs of inputs. The cost of preschool for children in the control group is discounted (including the cost of parental care), but no deadweight loss associated with extra cost is considered. Using a 3% discount rate, their total benefits per

child are $208,927 in 2014 dollars. The program benefits-cost ratio is estimated at 2.5:1. This is positive, but considerably worse than the estimate for PPP. However, the benefits considered by Barnett and Masse (2007) only include health benefits related to smoking, and do not include costs related to crime. Moreover, their earnings are measured at age 21, so it is very likely that these benefit estimates are underestimated.

4.3 Results on Later-life Outcomes

We present the results for later-life outcomes of ABC and PPP in Table 5. This table is based on Heckman et al. (2010a) and Campbell et al. (2014a). The results are consistent with the overall discussion in the paper: we evaluate PPP and ABC as very successful programs in boosting later life outcomes. To make interpretation easier, we reverse negative outcomes such as obesity. While these results are selected out of a large number of results that were reported in each of the papers, they achieve significance levels after controlling for size to account for multiple hypotheses tests among selected groups of outcomes. The results presented here are some of the most important impacts found for PPP in labor outcomes and some of the most important impacts found for ABC in health. We can see in the table that the increase in HS graduation in PPP was 56% for the females, and that the increase in current employment percentage was close to 30% for both males and females. For ABC, there are important reductions in adult abdominal obesity of around 20% for females and 30% for males. The differences in health indicators shown in this table can have very strong consequences in health later in the life cycle. The takeaway from this table is that these demonstration programs dramatically changed the lives of their beneficiaries.

Table 5: Selected Treatment Effects at Adulthood by Gender, PPP and ABC

	PPP			ABC	
	Female	Male		Female	Male
HS Graduation, age 19	0.56	0.02	**Abdominal Obesity, age 30**	0.198	0.294
Partial linear *pvalue*	**(0.000)**	0.416	Block *pvalue*	**(0.080)**	(0.137)
Step-down, Partial linear*pvalue*	**[0.000]**	0.583	Step-down *pvalue*	**[0.080]**	[0.218]
Monthly Earnings, age 27	0.64	1.01	**Obesity and Hypertension, , age 30**	-0.028	0.529
Partial linear *pvalue*	**(0.000)**	**(0.011)**	Block *pvalue*	(0.501)	**(0.016)**
Step-down, Partial linear*pvalue*	**[0.000]**	**[0.037]**	Step-down *pvalue*	[0.641]	**[0.016]**
Current Employment, age 27–40	0.28	0.29	**Framingham Risk Score, age 30**	0.331	3.253
Partial linear *pvalue*	**(0.042)**	**(0.011)**	Block *pvalue*	**(0.070)**	**(0.038)**
Step-down, Partial linear*pvalue*	**[0.094]**	**[0.024]**	Step-down *pvalue*	**[0.070]**	**[0.038]**

Note: initial sample sizes are: **PPP**: 123; **ABC**: 111. PPP results taken from Heckman et al. (2010a). One-sided $p-value$ based on the FreedmanLane procedure, using the linear covariates maternal employment, paternal presence, and StanfordBinet IQ, and restricting permutation orbits within strata formed by Socioeconomic Status index (SES) being above or below the sample median and permuting siblings as a block (in parenthesis). Adjusted for step-down –i.e., multiple hypothesis testing– [in brackets]. ABC results taken from Campbell et al. (2014a). One-sided block permutation $p-value$ (in parenthesis). Adjusted for step-down –i.e., multiple hypothesis testing– [in brackets]. **PPP** monthly earnings is in thousands of 2010 dollars at age 27. High school graduation (at age 19) and current employment (at age 27) are self-explanatory. **ABC** all at age 34. Abdominal obesity: waist-hip rario > 0.9 for males; > 0.85 for females. Obesity and hypertension: interaction of two variables. Obesity: ≥ 30 in Body-mass index. Hypertension: systolic blood pressure ≥ 140 & diastolic ≥ 90. Framingham risk score: is a sex-specific algorithm used to estimate the 10-year cardiovascular risk of an individual – it is based on basic health indicators and habits.

4.4 Cost-Benefit Analysis

Table 6 presents the results on costs and benefits from two of the studies that were discussed in this section, and for another important study yet to be discussed. This table shows PPP had outstanding benefits to cost relationships. Note that cost-benefit calculations on ABC do not account for crime; for PPP, they do not account for child care benefits; and for both ABC and PPP, they do not account for missing maternal earnings. The evidence in this table is, to the best of our knowledge, the best available on cost-benefit relationships of demonstration programs. It shows a high level of effectiveness for the two programs. Many other studies that do not have long-term data do cost-benefit analysis using forecasts. However, the forecasts are usually very rough, and have scarce theoretical or empirical support. We do not include them in this analysis.

Table 6: Costs and Benefits of PPP and ABC

	PPP	ABC
Program Cost	-$20,956	-$83,788
Program Benefits		
Child care		$36,448
Child abuse and neglect		
Earnings	$92,052	$49,541
Earnings of Future Generations		$7,553
Maternal Earnings		$90,721
Smoking/Health		$23,471
Education Costs	$5,104	$935
Welfare/AFDC	$4,364	$259
Crime	$78,800	
Total benefits	$180,319	$208,927
Net present value	$159,363	$125,139
Benefit-cost ratio	8.6:1	2.5:1
(s.e.)	(3.9)	

Note: dead-weight cost is zero, 3% rate of discount, all values are in 2014 dollars. Standard errors are obtained using bootstrapping. PPP estimates from Heckman et al. (2010b); ABC estimates from Barnett and Masse (2007). Education costs include K-12 and higher education costs.

4.5 The Carolina Approach to Responsive Education (CARE)

4.5.1 Program Description

Project CARE, created and implemented by the same organization as ABC, began immediately after ABC ended. It compared the efficacy of two service delivery models: one group received a similar center-based treatment as ABC, adding weekly home visits from the child's teacher to the family. In the other group, home visitors made weekly visits but center-based child care was not offered (Wasik et al., 1990). In both approaches, the frequency of the home visits decreased as children aged. In this paper, we only compare the high intensity group with the control group in our estimations for simplicity, and because we focus on demonstration programs with a center-based component. However, CARE's relatively small sample makes it hard to form conclusions.

The literature on CARE is not nearly as extensive as the one on ABC. Maybe the two most interesting studies are Wasik et al. (1990) and Campbell et al. (2008). The first

presents outcomes up to 54 months of age, focusing on IQ measurements. The patterns of the impacts are surprising: the scores of the intensive treatment group are consistently higher than the other two groups, but the control group is consistently higher than the home visit group. One possible reason is that the control group children attended other preschool centers slightly more than the home visit group. However, the differences are not significant, perhaps due to small sample size. The latter paper finds treatment impacts of the intensive demonstration program in college attendance when ABC and CARE samples are pooled.

4.6 Infant Health and Development Program (IHDP)

4.6.1 Program Description

IHDP was a multi-site randomized demonstration program designed as multiple replications of ABC implemented in 1985. It was implemented by schools of medicine and hospitals in 8 different sites across the United States. Maybe its most unique characteristic is that instead of targeting children based on socioeconomic status, it targeted premature infants ($<$ 37 weeks of gestational age) with low birth weight. The creators of the program distinguished two different strata from the beginning of the program: the lighter low birth weight (LLBW) group ($\leq$ 2000 grams) and the heavier low birth weight (HLBW) group (2001-2500 grams).

The demonstration program was similar to the intensive treatment of ABC and CARE: it included center-based care, weekly or bi-weekly home visits, and medical services. However, it only lasted for the first three years of life, not for five, as in ABC or CARE.

4.6.2 Evidence from the Literature

The literature on IHDP has usually obtained estimations separately for the two weight groups, obtaining stronger impacts for the HLBW group (Brooks-Gunn et al., 1994b; McCormick et al., 2006). As an example, the age 18 impacts of IHDP reported by McCormick et al. (2006) show negative but mild impacts of the demonstration program for the LLBW group (3 points in verbal IQ, 4 in reading achievement). Yet, there are positive effects of

the demonstration program for the HLBW group (5 points in verbal IQ, and in reading achievement, and 6 in math achievement).

In this paper, we present only one general estimate for the program. Moreover, we use all the children in the sample, including those who are not part of low-income families. Having middle class families is a unique feature of IHDP that sets it apart from all of the RCTs that we consider in this survey. As opposed to the previous evaluations, the pooled results we present, evidence little effectiveness of the program.

However, different studies have pointed out that the relatively smaller impacts of IHDP may be due to its focus on non-economically vulnerable populations. In the first of those studies, Brooks-Gunn et al. (1992) discuss impacts on IQ scores at three years of age. They find that children of mothers with a college degree did not increase their scores, while children of mothers with high school degrees or less did. They also study whether the impacts were different by birth weight, and they find that in several of the demographic groups, it did not matter, and in the one group that mattered, heavier birth weight babies benefited more from the program. These two results point to the selection mechanism of IHDP, operating towards lower treatment effects. A recent paper, Duncan and Sojourner (2013), focuses on the income level of mothers and finds similar results: for the HLBW sample, the program has a significant and strong impact in IQ at ages 3 and 5 for children from disadvantaged families. Disadvantaged children gain 1.32 and 0.87 standard deviations more than higher-income children, who gain 0.32 and -0.26 respectively at ages 3 and 5. For higher-income children these gains are non-significant, but for low-income children they are highly significant.

4.7 Early Training Project (ETP)

4.7.1 Program Description

ETP is the oldest of the demonstration programs we cover. It was implemented from 1962 to 1965 in Tennessee as a randomized experiment aimed for black, disadvantaged children. It was less intensive than PPP, relying strongly on the home visits, and having summer schools

for children, but not traditional preschool services for the rest of the year. Table 3 shows the characteristics of the program (Gray and Klaus, 1965; Gray, 1969; Klaus and Gray, 1968). Tables 4 shows the characteristics of participants.

Gray et al. (1982) evaluates ETP: for IQ, they find the usual pattern of very strong initial impacts (approximately one standard deviation for the experimental group with higher intensity) followed by fadeout of the impacts during elementary school. The program does not seem to have had a positive effect on employment. The experimental groups members were around 15% more likely than the control group members to attend college after finishing school, but the differences were not significant. Overall, the pattern of results seems similar but weaker than the pattern in the PPP. This might be because the program was less intense, including the lack of year-round center-based activities.

4.8 Results on Youth Outcomes

We present the results for youth outcomes in Table 7. We base the interpretation of these results on non-parametric one sided p-values. We present step-down p-values to provide further information for the interested reader. However, we do not use them as the basis of our interpretation, as our objective is not to test whether the outcomes are jointly significant. Rather, our objective is to test whether each outcome of interest is significant by itself. It is important to note that, in many cases, pooling the data for men and women reinforces power and, therefore, increases the significance of the results.

For PPP, the first year of post-treatment results for IQ align with F4-1 in Section 1. For both males and females, the results indicate a boost on IQ of more than half a standard deviation. The results are very similar for the rest of the demonstration programs. The effects for ABC and IHDP are also very sizable at the beginning and fade out in the long-term, though not completely. This further supports F4-1. CARE has much lower impacts even at the beginning. This could merely be a consequence of the lack of precision caused by CARE's small sample size. We only contrast the control branch with the full-treatment

branch. Thus, our treatment group has only 17 observations—which is further exacerbated by the fact that we estimate effects by gender. The effects for ETP do not go beyond half of a standard deviation but are sizable as well.

The evidence on non-cognitive skills is less clearly summarized, but let us further qualify this statement. Character is measured from batteries reflecting various behaviors (see Heckman and Kautz, 2012), and in many studies, the data available to measure character is sparse. Moreover, these batteries are not population-standardized and are not usually comparable across studies of programs. Nonetheless, the literature reports effects on character for PPP, ABC, and IHDP. For CARE there are no statistically significant effects, and we speculate that the reasons are the same as in the case of cognition. For ETP, the measures are unreliable, because there are relatively few measures, and they do not seem to be directed towards any particular skill.

Table 7: Treatment Effects on Skills by Gender, All the Programs

	PPP			**ABC**			**CARE**			**IHDP**			**ETP**		
	Female	Male	Pooled	Female	Male	Pooled	Female	Male	Pooled	Female	Male	Pooled	Female	Male	Pooled
IQ, 1st Yr Post-Treatment	10.029	8.157	10.444	10.405	7.648	12.997	2.068	-6.172	6.840	10.677	14.265	11.560	6.546	4.132	3.598
Non-Parametric *pvalue*	**(0.000)**	**(0.008)**	**(0.000)**	**(0.000)**	**(0.014)**	**(0.000)**	(0.372)	(0.782)	**(0.056)**	**(0.000)**	**(0.000)**	**(0.000)**	(0.177)	(0.303)	(0.235)
Step-Down, Non-Parametric *pvalue*	**[0.000]**	**[0.026]**	**[0.000]**	**[0.000]**	**[0.090]**	**[0.000]**	[0.832]	[0.96]	[0.252]	**[0.000]**	**[0.000]**	**[0.000]**	[0.225]	[0.375]	[0.333]
Achievement	0.299	0.250	0.342	0.435	0.223	0.674	-0.004	-1.019	0.398	0.045	0.215	0.094			
Non-Parametric *pvalue*	**(0.024)**	(0.172)	**(0.016)**	**(0.010)**	(0.194)	**(0.004)**	(0.502)	(0.932)	(0.102)	(0.260)	**(0.024)**	**(0.084)**			
Step-Down, Non-Parametric *pvalue*	**[0.096]**	[0.668]	**[0.058]**	**[0.030]**	[0.562]	**[0.002]**	[0.960]	[0.996]	[0.344]	[0.650]	[0.118]	[0.296]			
Conscientiousness	0.174	0.056	0.239	0.142	0.050	0.396	-0.363	-1.123	-0.028	-0.007	0.200	0.035			
Non-Parametric *pvalue*	**(0.056)**	(0.386)	**(0.054)**	(0.194)	(0.430)	**(0.040)**	(0.828)	(0.920)	(0.566)	(0.554)	**(0.026)**	(0.312)			
Step-Down, Non-Parametric *pvalue*	**[0.090]**	[0.774]	**[0.082]**	[0.686]	[0.964]	[0.180]	[0.998]	[0.996]	[0.918]	[0.828]	[0.158]	[0.652]			

Note: initial sample sizes are: **PPP**: 123; **ABC**: 111; **CARE**: 64; **IHDP**: 985; **ETP**: 88. Non-parametric $p-value$ (in parentheses) accounts for compromised randomization, small sample size, and item non-response. Step-down $p-value$ [in brackets] accounts for the same and for multiple hypotheses testing. We perform a step-down procedure for all of the outcomes listed in Appendix C, but only include relevant subset of outcomes in the table. When calculating each treatment effect for each program we control for three baseline characteristics: (i) mother works; (ii) father at home; (iii) number of siblings. In ABC we add birth-weight and gestational age as controls to proxy for initial health conditions, because two individuals initially assigned to control status were swapped to treatment status "at the extreme urging of protective service officials". In PPP we add SES-index and Stanford Binet IQ score at age 3, because the randomization was implemented conditional on those variables. For details of data sources and construction see Appendix C. IQ is standardized to have mean 100 and standard deviation 15. Achievement and conscientiousness to have mean 0 and standard deviation 15.

A methodology for evaluating demonstration programs and a summary of estimation issues Appendix B. Details on the data sources are provided in Appendix C.

4.9 Major Findings

F4-1 **Demonstration programs strongly boost IQ in the short-run for disadvantaged children. However, treatment effects often fade out. For very early demonstration programs (before age 3), impacts might persist into adulthood. Early cognition mediates later-life outcomes.** For the 5 US-based educational demonstration programs that have followed participants into adulthood, the short-term effects on IQ, as measured by different IQ tests, have an approximate magnitude between .3 and .7 standard deviation. The impacts fade out, especially for programs targeting children after age 3.[10] In some cases, the treatment effect persists—to a much smaller magnitude—and in other cases, it completely vanishes after only a few years (see Hojman, 2015). The finding that early life cognition mediates later-life outcomes (see below) opens an important question about the importance of this early increase in cognition for later life outcomes. Figure 9 illustrates IQ dynamics for the Perry Preschool Program (PPP) as an example of the fade-out in the demonstration programs we analyze.

F4-2 **Demonstration programs boost non-cognitive skills in the short run for disadvantaged children. Due to the lack of data and because of the difficulties associated with measuring personality traits, it is often hard to determine the real size of the impacts on these skills and whether they are affected in the long run. Early non-cognitive skills mediate later-life outcomes.** For many demonstration programs, we are able to document treatment effects on non-cognitive skills in the short run. These effects appear to persist.[11] Non-cognitive skills

[10] ABC, which began when the children were born, showed lasting treatment effects on IQ throughout adulthood.

[11] See Almlund et al. (2011) for discussions on relevance, dynamics, measurement, and interpretation of

are fundamental in mediating later-life outcomes (Heckman et al., 2013; Heckman and Mosso, 2014). They appear to be more important than cognitive skills as mediators of many adult treatment effects (García and Heckman, 2014). Figure 10 illustrates this for the case of the Carolina Abecedarian Project (ABC).

F4-3 **Demonstration programs are especially effective in boosting early and later life outcomes (directly or through early skills) for disadvantaged children.** This is a recurring finding in the literature. We note, however, that most demonstration programs are targeted towards the disadvantaged. When they are not, the evidence supports greater effectiveness for children from disadvantaged families.[12] There are two main reasons for this finding. Parents with better socio-economic status typically provide children with an enriched home environment.[13] In addition, the substitutes for center-based childcare of less affluent children are generally of lower quality.[14]

F4-4 **Demonstration programs stimulate parenting and parent-child interactions.** A main mechanism for the success of demonstration programs is parenting stimulation (see García and Heckman, 2014; Heckman and Mosso, 2014). Improved parental environments and practices enhance the development of children in a lasting way. Importantly, there are unique difficulties in parenting practice within disadvantaged families as Lareau (2011), Badev and Cunha (2013), and Cunha (2012) document and Table 1 illustrates.

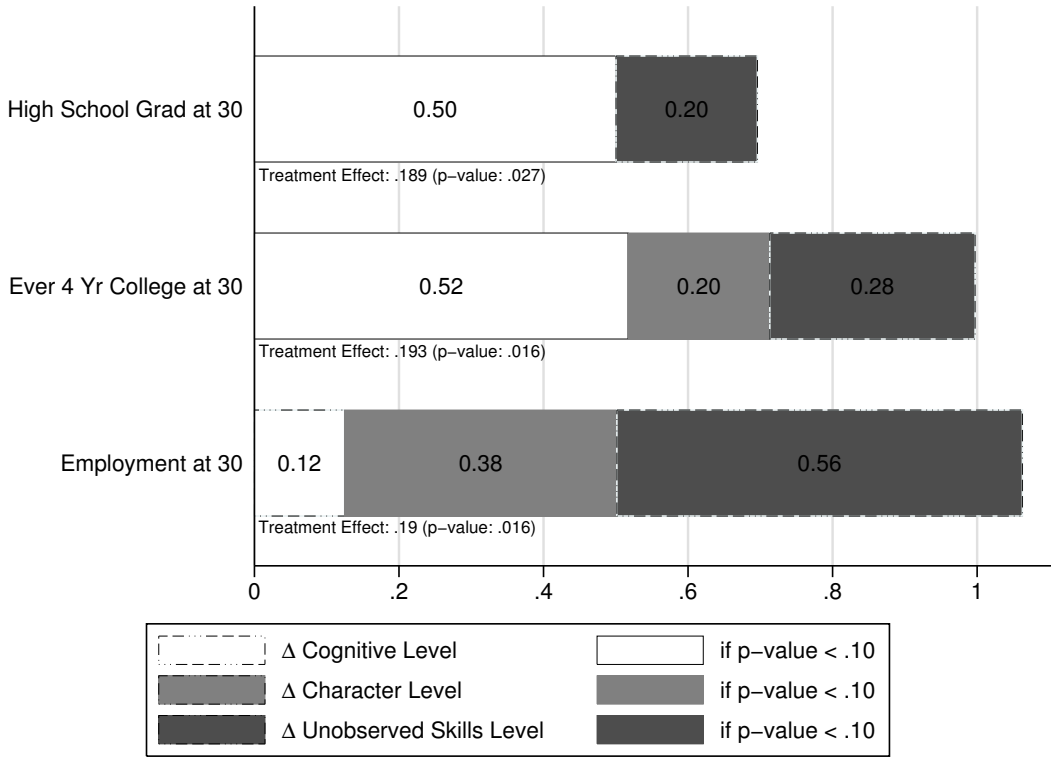
non-cognitive skills.

[12] An example of this comes from the Infant Health and Development Program (IHDP); see Section 4.6. This is the only demonstration program with long-term follow-up in which we find substantive variation across children's socioeconomic backgrounds. The effects of IHDP are much stronger for disadvantaged children (see Brooks-Gunn et al., 1992; Duncan and Sojourner, 2013; Heckman and Setzler, 2014).

[13] By design, the implementation of demonstration programs could crowd out parent-child interaction. It is important to grasp this and analyze the alternative to parent-child interaction. García et al. (2014) find that the reduction in parental-time interaction diminishes the positive total effects from a particular demonstration program on cognitive and non-cognitive skills.

[14] Bernal (2008a), Bernal and Keane (2010), and Bernal and Keane (2011) find that when parental-time interaction is substituted for informal childcare, there is a negative impact on children's cognition. However, they find no effect when the substitution is towards center-based childcare. See also Del Boca et al. (2013) and the discussion of the evidence in Heckman and Mosso (2014).

Figure 10: Cognitive and Character Skills as Education and Employment Mediators of Treatment Effects, ABC All



Source: García and Heckman (2014). Note: this figure shows how cognitive and character (non-cognitive) skills mediate treatment effects in the relevant outcome displayed on the y-axis. For the cognitive measure, we use achievement as measured by the Woodcock-Johnson test at age 21. For the character measure, we use Conscientiousness at age 15, based on a factor summarizing a set of items designed to measure personality traits. We make arbitrary choices on the items based on what we understand as conscientiousness. To avoid the arbitrary scale issues that tests and factors have, we use a variable representing 25 quantiles instead of either the score test or the factor itself. The bars display the components of three-folded Laspeyres decomposition exercises on each relevant outcome. To obtain them, we estimate two models, one for the treatment group and one for the control group. We first estimate the treatment effect and display below the bar with its respective asymptotic p-value to test if it is statistically different than zero. Then, we construct differences in mean skills, which we call Δ (Cognitive Level) and Δ (Character Level). The difference of the mean skills is scaled up by the control coefficient. Δ (Unobserved Skills Levels) corresponds to the difference in the intercepts in each model. We interpret this as a residual, i.e. the proportion of variance cognitive and character skills cannot explain. The length of the bars is not identical to one because we avoid displaying the components accounting for control variables.

F4-5 **The available cost-benefit analyses suggest that investing in demonstration programs is not only socially fair but economically efficient.** A comprehensive cost-benefit analysis (CBA) includes long-run outcomes to account for life-cycle gains in income, employment, health, etc. Unfortunately, few demonstration programs have long-term data follow-ups. There are two demonstration programs with long term follow-ups, PPP and ABC. The CBA of PPP is a long-term comprehensive study accounting for later life outcomes, with enhancements in earnings and reductions in

crime as its main components (see Heckman et al., 2010b). It accounts for dead-weight losses of taxation aimed to finance the program and nonetheless still finds an annual rate of return between 7% and 10%, which is above the long-term stock market rate of return on equity of 5.8% in the post World War II era before 2008 (see Heckman et al., 2010b). The CBA for ABC is not so wide in scope (Barnett and Masse, 2007). It is only based on earnings through age 21. Not surprisingly, the rate of return they find is rerlatively low, 2.5%. Recent evidence through age 35 suggests that health outcomes and crime reduction are major components of the gains of ABC (Campbell et al., 2014a). The CBA results are reported in 4.4.

5 Head Start and Other Large-Scale Programs

In this section, we discuss early education programs that have been implemented in a large scale (with either state-wide or nation-wide scope). We first discuss methodological problems associated to the evaluations of these programs, especially the problem of substitution bias. Then we argue that the best evaluations of HS show that it is not a failure. Rather, it has mildly positive results. Finally, we briefly discuss the evidence on other large-scale programs. In doing so, we underscore that these programs, HS in particular, many times present methodological challenges that make it hard or even impossible to evaluate them in a rigorous way. To illustrate, we discuss the only five RCTs of large-scale EEPs that, to the best of our knowledge, might be considered useful, and discuss that it might be impossible to draw interesting conclusions even from them. We describe them in Table 8.

5.1 Issues in the Evaluations of RCTs

We begin by pointing out what we consider the main issue in the evaluation of these programs. We state it as one of the main findings we document in this paper.

Table 8: RCTs of Public Programs Including an Early Education Component

	Even Start	**CCDP**	**Early Head Start**	**HSIS**	**TN-VPK**
Program yrs.	1999/2000	1990/1991-1994/1995	1995/1996-1998/1999	2002/2003	2009/2010-2010/2011
Delivery mode	Family Literacy: child and adult education	Case management	Home Visits, Center-Based and mixed	Home Visits, Center-Based and mixed	Center-based
Eligible population	Parent eligible for adult educ. & child age ≤ 8 (average star child age 3.2)	Under poverty line Child age ≤ 1 (or pregnant)	90% under poverty line Child age ≤ 1 (or pregnant)	90% under poverty line Child age 3 (3yc) or 4 (4yc)	Priority to free lunch-eligible children
Sample size	18 grantees, 463 families	21 projects, 3961 families	17 grantees, 3001 families	383 grantees, 4667 children	3025 full sample, 1078 ISS
Last follow-up	After 2 years (avg. age 5)	Age 5	5th Grade (approx. age 11)	3rd Grade (approx. age 9)	1st Grade (approx. age 7)
% Response rate	76% (less in analyses)	74% Treated 78% Control	54.4%	76% Treated 71% Control	Around 96% Treated in ISS: approx. 60% Control in ISS: approx. 53%
Attrition correction	Apparently not	Apparently not	Yes	Yes	No
Parameter	Intent to Treat	Intent to Treat	LATE. Participant defined as receiving 1+ home visits or attending 2+ weeks to a center	Intent to Treat	LATE. Participant defined as found in admin. data (full) or attended 20+ days (ISS)
Experience of treated group	Year 2: 72% in child educ.; 28% in parenting educ. Avg. family participated 8 (of 12) mos.	33% enrolled 5 years Center-based child care: Age <3 avg. 36.6 hrs./mo. Age 3-5 avg. 53.9 hrs./mo.	73.3% had 20+ hrs. of center care or 1+ home visit per week at least one of the 3 years	Approx. 83% had some HS on the year of evaluation	ISS: Avg. 149 day attended
Experience of control group	Year 2: 32% in child educ.; 17% in parenting educ.; 16% in Even Start	Center-based child care: Age <3 avg. 19.2 hrs./mo. Age 3-5 avg. 36.8 hrs./mo.	13.8% had 20+ hrs. of center care or 1+ home visit per week at least one of the 3 years	Approx. 40% center care Approx. 15% Head Start	49% stayed with parents 15% private child care 11% Head Start
Multilevel estimation	No	Yes: sites inversely weighted proportional to variance	Yes: sites equally weighted		Yes
Multiple hypotheses	No correction	No correction	No correction	Benjamini-Hochberg	No correction
Impacts: Number of Positive Significant Effects / Number of Variables Measured					
Cognitive	After 2 years: 0/8	Age 3 and 5: 0/4 (PPVT-R, TVIP, Kaufman)	Age 3: 4/4. PPVT (ES 0.12) Age 10: 0/10	HS Year, 4yc: 7/12, 3yc: 8/12 PPVT ES 4yc: .09, 3yc: .18 Age 9, 4yc: 1/11 Age 9, 3yc: 0/11 (1 negative)	ISS, PreK: 7/7 Woodcock Johnson ES .31 ISS, 1st Grade: 0/10 (1 neg.) Woodcock Johnson: No impact
Personality	After 2 years: 0/5	Age 3: 0/6 (Child Behavior Checklist) Age 5: 0/3 (Child Behavior Checklist)	Age 3: 4/9. CBCL-Aggressive (ES -0.1) Age 10: 1/11 Socio-Emotional Index (ES 0.1)	HS Year, 4yc: 0/9, 3yc: 2/9 Hyperactive (ES -0.21) Age 9, 4yc 2/19 (4 negative) Age 9, 3yc 1/19	ISS, PreK: 3/6 Cooper-Farran Social (ES .34) ISS, 1st grade: 0/6 Full sample, KG: less retention
Parenting/ environment	After 2 years: 0/9	Age 3: 0/8 (NCAST) Age 5: 0/4 (NCAST)	Age 3: 8/17 HOME (ES 0.11) Age 10: 0/9	HS Year, 4yc: 1/5, 3yc: 3/5 Age 9, 4yc: 1/15, 3yc: 1/15	
Parents' self-sufficiency	NA	0/23 (employment, income, welfare)	Age 3: 2/8. Training weeks (ES .18) Age 10: 0/5	NA	NA

Note: abbreviations used in this table: CBCL: Child Behavior Checklist; CCDP: Comprehensive Child Development Program; ES: Effect Size; HOME: Home Observation for Measurement of Environment instrument; HSIS: Head Start Impact Study; HS: Head Start; ISS: Intensive Subsample (in TN-VPK); KG: Kindergarten; LATE: Local Average Treatment Effect; NCAST: Parent-Child Interaction Scales. PPVT: Peabody Picture Vocabulary Test; TVIP: Test de Vocabulario en Imagenes Peabody; PreK: Prekindergarten; TN-VPK: Tennessee Voluntary Prekindergarten Program. 3yc and 4yc: 3-years cohort and 4-years cohort (in HSIS). The row "Multilevel Estimation" indicates whether the estimators used to calculate the impacts of the program accounted for the hierarchical structure of the data. Outcome Section: For all sets of outcomes and age, we note the number of outcomes that were significant at the 5% level (one tail), and the total number of outcomes measured (for example, a 4/9 entry in cognitive means 4 significant impacts out of 9 cognitive tests that were taken). For the TN-VPK intervention, the ISS sample are the children in the control and treatment group for whom informed consent was obtained for testing. On the other hand, "full sample" is the complete sample, for which results were obtained in administratively-recorded outcomes. See the description of the program for an analysis of how the evaluation on the ISS subsample might be considered non-experimental.

F5-1 **Substitution bias is pervasive in recent program evaluations.**[15] **As the quality of alternatives to any program increases, the *measured* treatment effects of an offer to participate in a program decreases.** We use the term *substitution bias* to describe the phenomenon in which control group participants of a program find a comparable substitute to the treatment. In this case, the estimate of the impact of the program against a world without any substitutes is difficult to obtain. Evaluations of programs in which there is substitution bias follow two different strategies: (i) Estimate the treatment effects under traditional methods and estimate the treatment effect of participating in the program against the next best alternative of non-participants, be it participating in an alternative preschool program or not. Clearly, the better the alternatives, the smaller the estimated impacts of the program; (ii) Apply methodologies that address the possibility of multiple choices (e.g., program, other preschool, or no preschool) and self-selection and, therefore, obtain the various counterfactuals of interest—program or other preschool and program or no preschool. While the second strategy delivers more information, it has important methodological challenges. In an early study, Heckman (1974) addresses this problem.[16] Although the validity of evaluations of demonstration programs is also threatened by this issue, substitution bias is fundamental in large-scale federally funded RCTs.[17] Firstly, they have been implemented more recently, with more alternatives available than ever in the past. Moreover, as they are less intensive, they may be closer in their impacts to alternatives. In Table 8, we show that this problem is common to all the RCTs of large-scale programs. Interestingly, HSIS, which is a RCT evaluating Head Start, has the highest rate of substitution bias (40%) (Feller et al., 2014; Torcasso, 2014).

Other relevant issues that are presented in Table 8 include: (i) treatment heterogeneity—e.g., in HSIS, multiple curricula were used and the quality of the programs greatly varied

[15] See Heckman et al. (1999, 2000) for discussions of the general problem of substitution bias.

[16] See also Burchinal et al. (1989); Heckman et al. (2000); Campbell et al. (2014b); Kline and Walters (2014); Feller et al. (2014); Torcasso (2014); Heckman and Vytlacil (2007); Pinto (2014).

[17] See Section 4, in which we document that ABC as en example in which substitution bias is relevant.

across sites (see Torcasso, 2014; Walters, 2014); (ii) attrition is high and potentially biases treatment effect estimates due to selection of unobserved characteristics; (iii) long-term follow-ups are not available and, therefore, the long run benefits of the programs are practically impossible to assess. In this section, we develop further on these issues and discuss the two large-scale RCTs with better evaluations (HSIS and TN-VPK).

5.2 Head Start Impact Study (HSIS)

Head Start (HS) is by far the largest ECI in the US. Accordingly, HSIS is the largest RCT of an ECI that exists. For ethical concerns, HSIS is a randomized one-year version of Head Start that exploits the fact that there is an excess of demand for HS. It includes a 3-year-old cohort and a 4-year-old cohort. It randomized treatment across 4,677 children through 383 grantee sites. The most current follow-up to date was at third grade, with an attrition rate of 25%.

The intervention was randomized at the site level. While the control group was not granted treatment in the participating Head Start site, nothing prevented control individuals from applying to other Head Start sites. In fact, 40% of the control group used center-based care. Moreover, 15% of the individuals in the control group attended some other Head Start center.

The main report of the intervention, Puma et al. (2010), reports intent to treat effects (ITT). For the 3- and 4-year-old at baseline cohorts, 7 and 8 out of 12 cognitive outcomes, respectively, were significant and positively impacted. These treatment effects fade out strongly by age 9. The results on character measures were much smaller; only a few of them were positive. Moreover, the program had very few impacts on parenting (see Table 8). Given that an substantial number of the control group attended HS itself, the ITT estimates in this case have no meaningful economic interpretation (i.e. they do not represent a counterfactual state in which there are no HS programs). All of the remaining papers using HSIS data in this section start by estimating the Local Average Treatment Effects (LATE) of participating

in HS where the counterfactual is not participating in HS, which is the effect of treatment compared to the next best preferred option (see Heckman and Vytlacil, 2007; Heckman and Urzua, 2010). However, this estimate does not address the problem of substitution bias, as most of the children using substitute treatments attended preschools other than HS. The studies find impacts on cognition of around .2 standard deviations after one year of the intervention, with a decrease to around .1 after two years of the program (see, for example Ludwig and Phillips, 2008). These results are much smaller in magnitude compared to any of the results in Section 4. As we explain below, this is possibly a consequence of (i) the availability of substitutes; (ii) children's heterogeneous backgrounds; (iii) a relatively short program duration of one year.

Bitler et al. (2014) explore the heterogeneity of treatment effects in HSIS. They estimate Quantile Treatment Effects (QTE) and QTE under endogeneity, respectively analogous to ITT and LATE. They find that the impacts of HSIS are stronger in the lower percentiles of the distribution of impacts. They also find impacts by baseline scores. Consistently with the literature, they find that the impacts of the program are stronger in the bottom of the distribution (although they find an unusual pattern of U-shaped distribution of impacts).

5.2.1 Analyses of HSIS that Address Substitution Bias

Two recent and preliminary studies address substitution bias in the HSIS data. These are Feller et al. (2014) and Kline and Walters (2014). Both papers go beyond standard ITT and LATE using a framework that allows for multiple treatments. Both papers rely heavily on normality assumptions to secure estimates. Only Kline and Walters (2014) analyze self-selection into the various options available to participants. Both papers model families as having three choices: (i) Head Start, h; (ii) other Preschool Program, p; and (iii) staying at Home, n. They identify 5 exhaustive and mutually exclusive groups: (i) Always-Head Start (11%); (ii) Always-Preschool (11%); (iii) Always-Home (12%); (iv) Preschool-

Compliers (20%); and (v) Home-Compliers (45%).[18] They discuss the identification problems related to having a single randomization to identify the impacts of three different choices. This literature addresses the issue of substitution bias and shows that, while the proportions of the different strata can be identified directly from the data, the treatment effects cannot be identified without additional assumptions.

Both papers share some conclusions: both find that HS has an impact that is very similar to the other available alternative center-based preschool settings, and both find that HS has significant effects on test scores for specific groups of children. These effects are moderate but this is meaningful, especially considering that the interventions last only one year and the follow-up is short. However, the magnitudes of their preferred impacts are quite different—between 0.2 standard deviations in Feller et al. (2014) and 0.47 standard deviations in Kline and Walters (2014). The magnitudes are not really comparable, as Kline and Walters (2014) obtain an Average Treatment Effects parameter, while Feller et al. (2014) obtain a LATE for the home compliers. The first paper finds a negative selection in gains in HSIS, so it is natural that after correcting for selection, the impacts are higher.

The results speak to F5-1 and F5-2 in this section. First, they document that, as has been established in other applied literature, it is fundamental to account for substitution possibilities when evaluating ECIs. This is especially important in the case of large scale programs—the wider the intervention, the more substitutes available for families. Second, contrary to conventional wisdom, it is hard to make the case that Head Start is a failure. We document this finding further by assessing non-randomized evaluations of Head Start in Section 5.3.

F5-2 **The available evidence suggests that Head Start is not a failure, despite claims to the contrary.**[19] Though randomization can be invaluable in the evaluation of EEPs, Head Start Impact Study (HSIS) provides a representative example of when

[18] These numbers are from Feller et al. (2014). The numbers in Kline and Walters (2014) are very similar.

[19] See, for example, Andrew J. Coulson (2010); Steve Aos and Roxanne Lieb and Jim Mayfiel and Marna Miller and Annie Pennucci (2004); Muhlhausen and Lips (2010); Zigler and Styfco (2010); Barnett (2011); Haskins (2010); Klein (2011); Burke and Muhlhausen (2013).

it is not enough. HSIS highlights the complexity of many social experiments, especially the problem of substitution bias. When substitution bias is addressed, the effects of Head Start on different measures of cognition and achievement appear to be higher as in Kline and Walters (2014). Careful non-experimental evaluations on Head Start using very different methodologies also report positive results of the programs in dimensions including some of the following childhood achievement, labor market, behavioral and health outcomes (see Currie and Thomas, 1995; Garces et al., 2002; Ludwig and Miller, 2007; Deming, 2009; Carneiro and Ginja, 2014). The literature generally shows that Head Start is a program with some important impacts, at least for children who would otherwise stay at home.

5.3 Non-randomized Evaluations of Head Start

Table 9: Selected Evidence on 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)	Perry Preschool (own calculations)	Abecedarian (own calculations)
Dataset	C-NLSY	PSID	Multiple	C-NLSY	C-NLSY	HSIS	HSIS	Own data	Own data
Subpopulation	Black	Black, mother educ. $\leq$ high school		Black	Males			Black, low child IQ at entry & SES	98% Black, low mother IQ
Years of birth	1979-1987	1966-1977	1960-1975	1979-1986	1977-1996	1998-1999	1998-1999	1959-1964	1972-1977
Impacts									
IQ/achievement, age 3-4	-	-	-	-	-	**0.180**	**0.473**	**0.696**	**0.866**
	-	-	-	-	-	(0.026)	(0.120)	(0.136)	(0.181)
IQ/achievement, age 5-6	**0.44**	-	-	**0.287**	-	**0.110**	-	**0.297**	**0.359**
	(0.129)	-	-	(0.095)	-	(0.23)	-	(0.122)	(0.153)
IQ/achievement, age 7-21	0.201	-	-	**0.127**	-	**0.100**	-	-0.109	**0.226**
	(NA)	-	-	(0.075)	-	(0.026)	-	(0.149)	(0.142)
Grade retention ever	0.218	-	-	**-0.107**	-	-	-	-	-
	(0.295)	-	-	(0.056)	-	-	-	-	-
High School grad. (no GED)	-	0.00	0.117	0.067	-	-	-	-	-
	-	(0.071)	(0.080)	(0.044)	-	-	-	-	-
Attended some college	-	0.031	**0.028**	**0.136**	-	-	-	-	-
	-	(0.067)	(0.019)	0.049	-	-	-	-	-
Earnings, age 23-40	-	0.051	-	-	-	-	-	$7,584	$7,249
	-	(0.357)	-	-	-	-	-	(6,299.82)	(6,480.17)
Idle (no work or study)	-	-	-	-0.030	-	-	-	**0.204**	**0.273**
	-	-	-	(0.053)	-	-	-	(0.093)	(0.100)
Ever booked/charged crime	-	**-0.126**	-	0.051	-	-	-	-	-
	-	(0.05)	-	0.050	-	-	-	-	-
Behavior Problem Index, age 12-13	-	-	-	-	-0.647	-	-	-	-
	-	-	-	-	(0.582)	-	-	-	-
CES Depression Scale, age 16-17	-	-	-	-	-0.552	-	-	-	-
	-	-	-	-	(0.489)	-	-	-	-
Head Start Related Mortality, ages 5-9	-	-	**-1.198**	-	-	-	-	-	-
	-	-	(0.796)	-	-	-	-	-	-

Note: impacts are in bold whenever they would be significant in a t-test at the 10% significance level. 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 age 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. 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. Earnings in Garces et al. (2002) are measured in logs. Feller et al. (2014) originally reported 95% posterior intervals of (0.13, 0.23) during the Head Start Program, (0.06, 0.15) during pre-K/Kindergarten, and (0.05, 0.15) during Kindergarten/first grade. 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.

Many researchers use diverse econometric methods and quasi-experimental designs to directly evaluate Head Start. We present here the studies we consider relatively strong from a methodological point of view. The findings are, in general, mildly positive for HS, which we reflect in our finding F5-2.

5.3.1 Fixed Effects Studies

Currie and Thomas (1995, 1999); Garces et al. (2002), and Deming (2009) use family fixed effects in their identification strategy. This relies on strong assumptions. They control for the effects of competing early childhood programs by conditioning on attendance to other preschools. Currie and Thomas (1995) find that there are gains in test scores of equal magnitudes for whites and blacks at the end of the program, the gains fade out during elementary school for blacks—which speaks to F4-1 in Section 1. Similarly, Deming (2009) finds no effect for whites in test scores. He finds effects for blacks which strongly fade out. Both papers find beneficial impacts in grade repetition, the first for white children and the second for black children.

Garces et al. (2002) and Deming (2009) measure treatment effects for adult outcomes. Both find impacts on high school completion and some college attendance. The first for whites and the second for blacks. While Garces et al. (2002) finds impacts in crime for blacks, Deming (2009) finds no impacts on crime at all. The first paper does not find significant impacts in earnings—possibly because the respondents were, on average, only 23 years old—and the second does not measure it.

The most puzzling pattern is the larger impacts of the program for whites in Currie and Thomas (1995) and Garces et al. (2002). This is inconsistent with F4-3 in Section 1. However, Currie and Thomas (2000) suggest that blacks went to lower-quality HS, which might imply that there is not really an inconsistency with our findings but a comparability problem within the populations.

5.3.2 Regression Discontinuity Designs

Carneiro and Ginja (2014) study the long-term impact of Head Start on health (frequent visits to the doctor, frequent use of medicines, ever having smoked, etc.) and behavior (grade repetition, school damage, special education etc.) outcomes. They exploit exogenous variation in program eligibility rules, which imply multiple discontinuities based on year, state, family size, and family structure to identify diverse margins enabling them to estimate causal effects of interest. The multiplicity in the eligibility thresholds distinguishes this paper from classic regression discontinuity designs. They do not focus on individuals located in a single threshold. Instead, they include various cohorts and, thus, provide effects informing causal impacts for different population groups.

This identification strategy enables them to identify the marginal effect of relaxing the eligibility requirement of Head Start. A somewhat puzzling empirical finding of the paper is that eligibility discontinuities shift the propensity to participate of Head Start only for boys. The interpretation of this is that the marginal individuals for which the effects are found are boys. The authors do not take a stand on a reason why this may be true.

The authors combine data from the National Longitudinal Survey of Young 1979 (NLSY79) and the children of the NLSY79, the cNLSY79. The former is a nationally representative survey of the individuals aged 14 to 22 years old in 1979. The latter surveys the children (or surveys mothers about their children). Matching these two samples, the authors use the household-level information to replicate the Head Start eligibility rules that applied to each of the individuals in the sample drawing from the NLSY79 and draw the individual specific outcomes from the cNLSY79. Importantly, it is hard to state the extent to which cNLSY79 is nationally representative, since it surveys the children of the NLSY79 without further sampling design. By sample design, the authors are able to assess impacts of Head Start for individuals receiving treatment in the 1980s and the early 1990s. In summary, the findings are the following. The program decreases behavior problems, chronic diseases, and obesity at ages 12 and 13, depression and obesity at ages 16 and 17, and crime at ages 20

and 21.

Ludwig and Miller (2007) exploit information on technical assistance on the implementation of HS to 300 poor counties offered by the Office of Economic Opportunity. The authors find an empirical discontinuity in the participation and funding rates, which is 50% to 100% higher in the 300 counties just above the cutoff for HS eligibility. The authors point out that the discontinuity in other federal social spending is small and not significant and, therefore, argue that the discontinuities on which their estimations are based are not only exogenous but solely produced by the technical assistance that HS provides. The main estimates are based on a regression discontinuity design comparing treatment and control counties near the cutoff. That is, the "treatment group" consists of individuals in the 300 counties who received technical assistance and the "control group" are individuals in counties with characteristics very similar to the treatment counties but did not receive any treatment whatsoever. The authors find reductions in mortality and evidence on schooling improvement (it is suggestive due to data limitations).

5.4 Tennessee Voluntary Prekinder Program (TN-VPK)

This is the only state-funded and state-implemented EEP that has been studied with an RCT. A description of this program is in Table 8. This intervention has one important problem: program implementers requested parental consent after performing the randomization. This affected consent. The sample that gave consent is referred to as the Intensive Substudy (ISS). They were assessed by the research team with a battery of achievement measures. For cohort 1, 46% of the treatment group parents and 32% of the control groups parents gave their consent to be part of the ISS and for cohort 2: 74% and 68%. This creates a major issue of selective attrition.

There was a positive impact on school retention, an outcome that was measured both for the whole sample (4.0% retention for the treatment group compared with 8.0% for the controls) and for the ISS (4.1% and 6.2%, respectively). The ISS showed positive treatment

effects during the preschool year in all of the direct assessments of academic skills. The effect size for the Woodcock Johnson Test for achievement was 0.24. However, at the end of first grade, there were no effects on cognitive skills and no differences for personality outcomes. The authors of this report were very cautious on discussing these results, stressing that more time was necessary to obtain conclusions (see Lipsey et al. (2013a) and Lipsey et al. (2013b)).

5.5 The Chicago Child-Parent Center (CPC)

CPC began in 1967 and targets disadvantaged children. It is the second oldest federally funded preschool program after Head Start. Assignment is not random, but follow-ups are available both for participants and non-participants up to age 26. Its evaluation compares children attending preschool and kindergarten to individuals who did not go to preschool but did go to kindergarten.

Gensowski et al. (2014) reanalyze CPC. They test and control for selection through a standard control function approach (see Heckman, 1979). They incorporate the information on the quality of the children's neighborhood and their access to the program. Similar to Reynolds et al. (2011), Gensowski et al. (2014) find that the program increased men's educational attainment by over a year and significantly improved the economic outcomes for both men and women. However, the authors do not find statistically significant impacts on criminal behavior. Gensowski et al. (2014) also find that, contrary to the literature, the CPC was less effective for children from severely disadvantaged backgrounds.

6 Universal Preschool & Childcare Policies: the Available Evidence

We devote this section to universal childcare and universal preschool programs because of their importance in policy debates. By universal, we mean programs that are available to all children in an area, though the programs we discuss in this section may have regional,

state-wide, or national scope. Unfortunately, the evidence on universal childcare for the U.S. is scarce. Therefore, we focus on states in the US with state-wide preschool programs and some nation-wide policies implemented in Canada and Norway. Although it is difficult to extrapolate from policies in other countries to recommend policy for the US, we think that the patterns are driven by economic forces that hold independently of the location of the policy—e.g., behavioral responses in labor force participation after childcare policy. We divide our findings into two categories: (i) parental labor force participation and (ii) child skill development. Parental labor force participation is relevant to early child development, as it affects the type of early care environment, time spent with parents, and financial resources that parents can invest in their children. Although they can both potentially increase parental labor supply, there is a major difference between universal childcare programs and universal preschool programs. The former are heterogeneous in quality and may not have a well-defined educational component, while the latter are all EEPs. However, despite this major difference, there is one common finding: *universal policies are not always universally beneficial*. The available evidence shows stronger positive impacts for disadvantaged children, while impacts on relatively more advantaged children is weak or even negative. We summarize the main lessons of this section in the next two findings.

F6-1 **Universal childcare policies have heterogeneous effects across children belonging to families from different socio-economic statuses.** Our interpretation of the results is that disadvantaged children benefit the most from universal policies. Potentially, this is due to the high quality of these programs compared to the children's alternatives. For children from more advantaged backgrounds, universal childcare can mean losing valuable interaction with high human capital parents or being placed in a program that may be of lower quality than a more expensive alternative. For children from disadvantaged backgrounds (e.g. single parent households as described in Section 3), universal childcare implies avoiding situations in which they are taken care of in disadvantaged, informal childcare arrangements or environments. This is related to

Finding F4-3 (see Gormley and Gayer (2005); Havnes and Mogstad (2011b); Bitler et al. (2014); Fitzpatrick (2008); Cascio and Schanzenbach (2013); Baker (2011)).

F6-2 **Impacts of programs are tied inextricably to program quality. This is especially relevant in the few examples of universal childcare in the US.** Assessments of universal and generic childcare programs such as Bernal and Keane (2010) and Bernal and Keane (2011) find negative effects to time spent in non-maternal care; however, this is only true for children attending informal care settings. The intuition is that the program must be of a higher quality than the available alternatives—even a well-intentioned subsidy may have negative impacts if it means, for example, that parents substitute the time they spend caring for children with time the children spend in a low-quality informal setting. This is the same intuition behind Finding F6-1 (see Havnes and Mogstad (2011b), Havnes and Mogstad (2014), Baker et al. (2008)).

6.1 Universal Childcare

Three main sources provide evidence on the impacts of attendance to childcare for children of heterogeneous backgrounds (i) a national reform in Norway; (ii) a local reform in Quebec; (iii) an analysis of all types of childcare used in the NLSY79. These programs and their impacts on maternal labor force participation and child development are displayed in Table 10. The Norway program provided a subsidy to formal childcare centers serving children aged 3–6, effectively making childcare cheaper for families. It had clear conditions for quality with low staff-to-child ratios and standards for staff quality. The Quebec program subsidized families directly, allowing them to use both formal and informal care. It had a substantial portion of non-educational centers, the quality of which is hard to characterize. This may result in differences in mean outcomes between them. As discussed below, the impacts for the Norway program were negative for well-off children, and positive for disadvantaged children. While this evidence is not available for the Quebec program, its impacts were negative on average, and it operated mostly on better-off families, which is suggestive evidence to support claims

F6-1 and F6-2.

Havnes and Mogstad (2011b) discuss the reform to childcare in Norway. It included federal provisions on educational content, group size, staff skill composition, and physical environment. The reform also increased the level of federal subsidies. The authors limit the evaluation to children of married mothers. They find positive effects of the reform on both education and labor market attachment, as well as welfare dependency. They also find an impact on delay in child bearing. Most of the effect on education comes from children with low-educated mothers, and most of the effect on labor market attachment and earnings is for girls. Havnes and Mogstad (2014) expand the previous analysis, aiming to explore the heterogeneity of the treatment effects of the reform. They find that the Quantile Treatment Effects in earnings are positive for most of the distribution of earnings but strongly negative for the top of the distribution. They also study how the impacts of the reform differ by family income. Their estimates suggest that "upperclass 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 (2011a) find very few effects of a universal preschool program on labor supply. As a consequence, the interpretation of the effects on children is that the expansion of preschool moved children from informal care, rather than from parental care, into formal care of relatively high quality.

Baker et al. (2008) present an evaluation of a universal childcare program in Quebec. The subsidized childcare is of heterogeneous quality and tends to be informal for younger children. The subsidy impacted especially children of higher-income families (low-income families were already being subsidized). Focusing on two-parents households, they find positive impacts on mother's labor supply (employment increased by 14.5% of the baseline level). However, impacts on children are negative, including aggression, motor, social skills (with impacts of around a whole standard deviation) and illness (a 237-451% impact in communicable respiratory illnesses for treated individuals). They also find that program led to more hostile parenting, worse parental health and lower quality parental relationships.

This paper is unique in covering all the most important effects of a childcare expansion. A later paper, (Kottelenberg and Lehrer, 2013), refine these estimates and show positive effects for the most disadvantaged children. This mirrors the case for universal preschool, for which some studies found that only those children at the tails benefited.

Bernal (2008a), Bernal and Keane (2010), and Bernal and Keane (2011) use the NLSY79 to examine childcare decisions in the US and their impacts on parental labor force participation and child development. They analyze the bredth of childcare options in the US, including formal and informal care options. They use three different methodologies to assess the impact of local child-care on cognitive and non-cognitive development: (i) a structural approach—using a fully structural model; (ii) a quasi-structural approach—using joint estimation of the reduced form decision rules and structural production functions; (iii) instrumental variables approach. The first paper uses a sample of married women. The last two use a sample of single mothers and exploit exogenous changes in the welfare structure as variation affecting the probability of the child being in child-care. The three papers show child-care has negative impacts on cognition, with an effect of 0.13-0.14 standard deviations. Importantly, Bernal and Keane (2011) show that this effect is mainly driven by informal child-care. While these studies do not analyze proper education programs, they inform F6-2—the effects of a childcare program are strongly affected by its quality.

Table 10: Comparing Heterogeneous Impacts of Childcare

	Norway	Bernal and Keane (2010) & Bernal and Keane (2011)	Quebec
Program Type	Subsidy to childcare centers (effectively making childcare cheaper for families).	Self-reported hours attending childcare.	Direct subsidies and tax credits on the basis of family income.
Target Population	Any child aged 3-6.	N/A	Children aged 0-4.
Program Quality	Eligible programs met regulations maximum price to be paid by the parents, educational content and activities, group size (staff-child ratio 1:8), staff skill composition (college-educated) and physical environment.	Highly heterogeneous. Studies consider quantity of childcare hours used.	2/3 of staff must be college educated and receive professional development. Staff-child ratio 1:10. Quality of care received did not change (relative to quality in the rest of Canada) as a result of the policy.
Treatment Heterogeneity	Homogeneous with respect to observable characteristics such as group size, staff-child ratio, and employee training.	High heterogeneity.	High heterogeneity.
Formal/Informal Care	Formal.	Both.	Formal, but including some home-based care.
Counterfactual	Children placed on waitlist; may have received informal care.	Maternal childcare.	Other formal or center-based care; pre-school; informal care.
Impacts on Child	No effect on test scores for males at age 18-20. Positive impact on earnings ($1,207 USD 2014) for low-income, but negative for high-income. Positive effects on yrs. of schooling driven by low- and mid-income.	Heterogeneous negative impacts on cognitive ability based on child's baseline skill and mother's education for informal care. No impact for formal care.	Positive effects for non-cognitive outcomes. Impacts suggest better outcomes for more disadvantaged groups.
Impacts on Maternal Labor Supply	No net impact on labor supply. Impact was mostly in crowding out informal care.	NA	Increased labor supply 14.5% of the baseline level.

Sources: **Norway**: Havnes and Mogstad (2011b) & Havnes and Mogstad (2011a). **Quebec**: Baker et al. (2008) and Kottelenberg and Lehrer (2013). Note: this table compares the characteristics of some large-scale childcare programs and thier impacts of parental labor force participation and child development.

6.2 Universal Preschool

Two recent studies analyze the implementation of universal preschool in Oklahoma and Georgia, Fitzpatrick (2008) and Cascio and Schanzenbach (2013). Using a difference-in-differences framework, Fitzpatrick (2008) finds positive effects of universal preschool in Georgia on academic achievement, as measured by math and reading test scores and staying on grade. Notably, Fitzpatrick (2008) finds that the effects of universal preschool were significant mainly for economically disadvantaged children in rural areas.

Similarly, using a difference-in-differences framework with state and cohort fixed effects, Cascio and Schanzenbach (2013) found positive effects on 4th grade test scores for students eligible for free or reduced price lunch in Georgia and Oklahoma. These effects lasted through eighth grade. However, they found negative effects on math scores for students not eligible for free or reduced price lunch. They posit that, based on the weaker evidence from targeted programs, positive effects of universal preschool might be mediated through peer effects—that the presence of high-quality students enriches the classroom environment, leading to increased performance for children on the margin. However, families of high-quality students may also benefit from needing to spend less money on care and education, freeing up resources that they may invest otherwise in their children.

6.2.1 Regression Discontinuity Designs to Evaluate State Preschool

A series of papers have analyzed the impacts of universal state preschool programs based on a regression discontinuity design on minimum entering age. They compare outcomes of children who will get the program, but have not yet received it, with a group of children that are slightly older and have finished their first preschool year. Unfortunately, these papers can only measure the effects on academic achievement of a single year of preschool, and they can only do so in a horizon of one year, as later the control group enters the program.

The best known example is the program at Tulsa, Oklahoma (see Gormley and Gayer (2005); Gormley Jr et al. (2005); Hill et al. (2012)). The program was of high quality, with

preschool teachers all having college degrees and certificates in early childhood and earning the same salaries as elementary school teachers. Teacher-to-child ratios are set at 1:10. The impacts on the this study are substantial, with an overall impact of almost 0.4 standard deviations in cognitive scores. Consistent with F6-1 and F4-3, the impacts are stronger on blacks, and much stronger on Hispanics, attending full-time programs. Interestingly, even if this paper is taken as evidence to support the case for universal preschool, the estimated effects on white children are non-significant (and negative in magnitude for the ones in full-day programs).

As discussed in this chapter, fadeout in cognitive impacts is pervasive in preschool programs, so it is likely that most, or all, of the impacts in this dimension will dissipate in the following years. It is unclear if the magnitude of the initial impacts on achievement are a valid proxy of the overall long-term impacts of the program, but if they are, Tulsa is a good example of the positive effects that a state-implemented high-quality program can have on minority children.

7 Structural Models of Parental Investment and Skill Formation

In this section, we present two simple models related to (i) educational programs and parental investment; (ii) parental investment decisions and skill development. To build intuition, we derive two separate cases. First, we model a static parental time allocation decision in which our objective is to understand what happens when the quality of publicly available EEPs increases. This seems to be in line with the findings noted throughout this paper, which are derived, mostly, as insights from offering high quality EEPs. Although this doesn't seem to take into account the extensive margin decision between formal and informal childcare, (i) around 80% of children attend some form of childcare (see Section 3); (ii) it is easy to imagine that the transition from informal to formal childcare happens in an intensive margin

of quality. Second, we draw on Heckman and Mosso (2014) and analyze the abstract process in which skills develop as an interactive process between investment and skills themselves. The objective of this is to understand what the dynamics of skill formation are and how they interact with parental investment. After presenting these two models, we summarize the evidence from structural models of parental investment to understand lessons from policy experiments related to investment in childcare and similar dimensions. These models contain pieces of the two models we derive first, since they account for investment decisions and child skill formation.

7.1 Parental Time Allocation and Early Education Quality

Consider a three person family which pools consumption, c. One of the parents always works and derives no utility from leisure. For simplicity, we refer to this parent as the father. His total labor income is y. Alternatively, we can think of a household with a single parent and an exogenous endowment y. The other parent, the mother, allocates her time to three activities: work, w; child investment, I, or leisure, l. A single uni-dimensional skill, θ, represents the child's human capital. Initially, an early education program with quality q is publicly provided. q is expressed in terms of its cost. Thus, it costs q to provide a program of quality q. The government pays for it by imposing a lump sum tax, τ. No other public goods are provided and no other economic activity is taxed. Therefore, the government decides to set $\tau = q$. The family utility function is

$$U = u_c(c) + u_l(l) + au_\theta(\theta) \tag{1}$$

where $a > 0$ measures the degree of parental altruism (see Becker and Tomes, 1979, 1986). We assume that the sub-utility functions in (1) are strictly increasing and concave in their respective single arguments. That is, $u_i' \equiv \frac{\partial u_i}{\partial i} > 0, u_{ii}'' \equiv \frac{\partial^2 u_i}{\partial i^2} < 0$ for $i = c, l, \theta$. The child's skill production is

$$\theta = f\left(eI, q\right) \tag{2}$$

where e represents the maternal efficiency of investing I in the child's human capital. $f(\cdot,\cdot)$ is strictly increasing and concave in its two arguments. $f_i \equiv \frac{\partial f(\cdot,\cdot)}{\partial i} > 0, f_{ii} \equiv \frac{\partial^2 f(\cdot,\cdot)}{\partial i^2} < 0$ for $i = eI, q$. Normalizing the mother's wage to 1 and giving her a one unit time endowment, the family budget constraint is

$$c + q = y + 1 - l - I. \tag{3}$$

Thus, the problem of the family is to maximize (1) subject to (3). We ask three questions: what happens with (i) investment on the child, (ii) maternal labor supply; and (iii) maternal leisure and labor supply when quality increases. The comparative statics follow directly from the first order conditions for an interior solution and are the following

$$\frac{dI}{dq} \begin{cases} < 0, & \text{if } f_{12} \leq 0 \\ > 0, & \text{if } f_{12}u''_\theta + f_1 f_2 u''_\theta > \Xi, \end{cases} \tag{4}$$

where $\Xi \equiv -\frac{u''_l u''_\theta}{ae(u''_c + u''_l)}$.

Intuitively, if the program and maternal investment are substitutes, an increase in the program's quality implies that maternal investment decreases. If the degree of complementarity is high enough, however, the mother invests more in the child. Complementarity needs to be high enough because there are two effects. On one hand, investing in the child becomes more productive because q acts as a production shifter when it increases exogenously. On the other hand, since q is taxed away, increasing it implies a reduction in consumption. Complementarity needs to be strong between investment and q to overcome the utility loss in consumption and make the mother decide to investment more on the child. Now, maternal labor supply increases after q rises under the following condition

$$\frac{dw}{dg} > 0 \tag{5}$$

if $ae^2u_c''f_{11} + f_1u_\theta''ae(eu_c''f_1 + u_l''f_2) + u_l''u_c'' > -aeu_l''u_\theta'f_{12}$. Otherwise, labor supply decreases.

Interestingly, it is possible to allow for a situation in which both maternal investment and labor supply increase. Naturally, it has to be the case that leisure diminishes while both investment and labor force increase. Formally,

$$\frac{dl}{dg} < 0 \tag{6}$$

if $ef_{11}u_\theta' + f_1u_\theta''(ef_1 - f_2) < f_{12}u_\theta'$. Otherwise, leisure decreases.

From this model, we learn that the complementarity between the program's quality and parental efficiency in investing is fundamental to understand whether a quality increase crowds out maternal investment. Also, we learn that it is not necessarily the case that an increase in quality, which is costly for the family, decreases investment. Instead, it is possible to have a scenario in which both maternal investment and maternal supply increase. The mother, then, needs to allocate less time to leisure. We expect to see this case if program's quality and maternal investment are complements to a great degree.

7.2 The Dynamics of Skill Formation

We draw on Heckman and Mosso (2014) and generalize the production of skills in time t, θ_t, to the following process

$$\theta_{t+1} = f_t\left(\underbrace{\theta_t}_{\text{self-productivy}}, \underbrace{I_t}_{\text{investment}}, \underbrace{\theta_{p,t}}_{\text{parental skills}}\right). \tag{7}$$

In general, both θ_t and $\theta_{p,t}$ may be vector-valued. The concepts we define here are easy to generalize to that case but we keep the scalar case for clarity. We say that f_t, the function

describing the law of motion for skills, exhibits the self-productivity property if

$$\frac{\partial \theta_{t+1}}{\partial \theta_t} > 0. \tag{8}$$

Self-productivity encompasses two ideas. Firstly, investments in skills do not fully depreciate. Secondly, the skills stocks and either investment or parental skills may be complements. If skills are vector-valued, different types of skills may be complements as well. This leads to the formalization of a crucial concept, dynamic complementarity:

$$\frac{\partial^2 \theta_{t+1}}{\partial \theta_t \partial I_t} > 0. \tag{9}$$

Recent studies claiming that investing early in skills is not only equitable but efficient actually rely on dynamic substitution at early stages. This is represented by,

$$\frac{\partial^2 \theta_{t+1}}{\partial \theta_t \partial I_t} < 0. \tag{10}$$

for $t < t^*$ and where t^* is a "sensitive period" in which the stock of skills and investments switch from substitutes to complements. Cunha and Heckman (2007) and Cunha et al. (2010) analyze and estimate the technology of skill formation and find that

$$\frac{\partial^2 \theta_{t+1}}{\partial \theta_t \partial I_t} \tag{11}$$

is an increasing function of time. This leads to two fundamental aspects of skill formation: (i) investments in relatively more skilled individuals become more productive as they age; and (ii) complementarity increases over the life cycle. Together, these two imply that later life remediation investment is less efficient than early life prevention and investment.

7.3 Evidence

Table 11 is taken from summarizes structural models of parental investments, reviewing their characteristics. Estimation results using NLSY79 and CNLSY support the conclusions from the beginning of this paper that complementarity of inputs increases with developmental stage (as does self-productivity). Interestingly, the correlation between parental investment and child skill also increase with developmental stage, though its impact is found to be largest in the earliest stages (see Cunha and Heckman (2007); Cunha (2007)).[20] Moreover, the parent's non-cognitive skill in more relevant than cognitive skill in the child's cognitive development. Studies using PSID corroborate that both parental time and good investment is crucial for child development (even increasing the probability of high school and college graduation), though goods become more important in later years (see Del Boca et al. (2014b); Gayle et al. (2013)).

In general, policy experiments indicate that in increase in early investments for children from low-income families not only boosts early achievement, but also results in an increase in college enrollment and could also decrease criminal behavior (see Cunha and Heckman (2007); Cunha (2007); Caucutt and Lochner (2012)). Subsidizing parents at these early stages, either by increasing borrowing limits or directly subsidizing investment, has similar results and also shows a small increase in entry wages in some cases (see Caucutt and Lochner (2012)).

[20] However, cognitive skill is shown to be a substitute for parental investment in early stages (Cunha, 2007).

Table 11: Structural Models of Parental Investments ("✓" means present; "X" means absent)

	Cunha and Heckman (2007)	Cunha (2007)	Caucutt and Lochner (2012)	Del Boca et al. (2014b)	Gayle et al. (2013)	Cunha et al. (2013)	Bernal (2008b)	Lee and Seshadri (2014)	Restuccia and Urrutia (2004)
OLG Model	✓	✓	✓	X	✓	X	X	✓	✓
Dynastic Links	A,B,C	A,B,C	B,C	X	A,C	X	X	A,B,C	A,C
Explicit Models of Parental Preferences, Altruism (a) or Paternalism (p)	✓(a)	✓(a)	✓(a)	✓(p)	✓(p)	✓(p)	✓(p)	✓(a)	✓(a)
Model Estimated	X	✓	✓	✓	✓	✓	✓	✓	X
Parental Goods Investment	✓	✓	✓	✓	✓	✓	✓	✓	✓
Parental Time Investment	X	X	X	✓	✓	X	✓	✓	X
Technology Depends on Parental Skill	✓	✓	X	X	✓	✓	✓	✓	X
Self-productivity	✓	✓	✓	✓	X	✓	✓	✓	✓
Parental Learning About Technology	X	X	X	X	X	✓	X	X	X
Bequests	✓	✓	X	X	X	X	X	✓	X
Intragenerational Borrowing	✓[E]	✓[D]	✓[E]	X	X	X	X	✓[D]	X
Multiple Skills of Children	X	X	X	X	X	X	X	X	X
Multichild Families: Allow for Preferential Treatment Across Children	X	X	X	✓	✓	X	X	X	X
Endogenous Fertility Decisions	X	X	X	X	✓	X	X	X	X
Multiple Parents	X	X	X	✓	✓	X	X	X	X
Endogenous Mating Decisions	X	X	X	X	✓	X	X	X	X

Source: Heckman and Mosso (2014) Note: A: Through parental skills, B: Through asset transfers, C: Through genes (initial conditions), [D]Natural borrowing limit, [E]Limits can be more stringent than natural limit.

8 Conclusion and Next Steps

We set out with the goal of reviewing the state of knowledge on the impacts of childcare in the US, focusing mainly on what we know about the impacts of EEPs. In particular, we aimed to evaluate the case for public provision of childcare and whether childcare policies should be targeted or universal. In doing so, we document 13 findings about EEPs in the US. We presented them in Section 1 and developed them throughout the paper.

The needs of families with working parents (especially in conjunction with the rise of single parent households in the US) has driven the increase in use of childcare, both formal and informal. However, there are clear gradients in the types and quality of childcare used by family structure, maternal employment status, and maternal education. These patterns seem to indicate that disadvantaged children receive lower quality care at a young age. At the same time, public preschool and other center-based programs provide a major source source of high-quality care, the former overtaking the latter in the last few decades. High-quality programs, especially small-scale, targeted interventions, have been shown to boost later-life outcomes by developing early-life skills. Furthermore, available cost-benefit analyses indicate that interventions are economically efficient. This evidence indicates that EEPs are very beneficial for child development. However, the evidence from large-scale interventions shows that their impacts are not only tied to some objective measure of program quality but to the quality of the program *relative* to the participating child's alternatives. That is, disadvantaged children benefit the most, while the impacts on relatively more advantaged children seem to be weak or even negative.

This would seem to indicate that means tested public childcare provision might be most effective. However, there is a much left to understand about the effect of large-scale programs on relatively advantaged children. This requires further research on how parents use the resources freed up by childcare subsidies. In short, the debate on early childcare policy is far from resolved. We still do not have a firm understanding of what aspects of high-quality, formal childcare cause positive impacts, nor do we fully understand how to

effectively implement them. Further research must be done to account for (i) the lack of information on parenting and child quality; (ii) the effects of subsidizing childcare on parental time and resource allocation; (iii) the characteristics of EEPs that drive their impacts.

References

Administration for Children and Families (2008). History of head start. http://www.acf.hhs.gov/programs/ohs/about/historyofheadstart. Accessed: 2014/09/02.

Administration for Children and Families (2009). Head Start Program Performance Standards. Available at: https://eclkc.ohs.acf.hhs.gov/hslc/standards/hspps.

Administration for Children and Families (2013). CCDF Fact Sheet. http://www.acf.hhs.gov/programs/occ/factsheetocc. Accessed: 2014/09/12.

Almlund, M., A. Duckworth, J. J. Heckman, and T. Kautz (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, and L. Wößmann (Eds.), *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Amsterdam: Elsevier.

Andrew J. Coulson (2010). Head start: a tragic waste of money. http://www.cato.org/publications/commentary/head-start-tragic-waste-money. Accessed: 2014/11/21.

Badev, A. and F. Cunha (2013). Maternal learning about the technology of skill formation. University of Pennsylvania.

Baker, M. (2011). Innis lecture: Universal early childhood interventions: what is the evidence base? *Canadian Journal of Economics/Revue canadienne d'économique 44*(4), 1069–1105.

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

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

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

Barnett, W. S. and D. J. Yarosz (2007). Who goes to preschool and why does it matter? *Preschool Policy Matters 7*.

Becker, G. S. and N. Tomes (1979, December). An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy 87*(6), 1153–1189.

Becker, G. S. and N. Tomes (1986, July). Human capital and the rise and fall of families. *Journal of Labor Economics 4*(3, Part 2), S1–S39.

Bernal, R. (2008a). The Effect of Maternal Employment and Child Care on Children's Cognitive Development. *International Economic Review 49*(4), 1173–1209.

Bernal, R. (2008b). 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 (2010). Quasi-structural Estimation of a Model of Childcare Choices and Child Cognitive Ability Production. *Journal of Econometrics 156*(1), 164–189.

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.

Bitler, M., H. Hoynes, and T. Domina (2014). Experimental Evidence on Distributional Effects of Head Start.

Blau, D. (2003). Child Care Subsidy Programs. In *Means-tested transfer programs in the United States*, pp. 443–516. University of Chicago Press.

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

Brooks-Gunn, J., R. T. Gross, H. C. Kraemer, D. Spiker, and S. Shapiro (1992). Enhancing the cognitive outcomes of low birth weight, premature infants: for whom is the intervention most effective? *Pediatrics 89*(6), 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. (1994a). Early intervention in low-birth-weight premature infants: results through age 5 years from the infant health and development program. *Jama 272*(16), 1257–1262.

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. (1994b). Early Intervention in Low-birth-weight Premature Infants: Results through Age 5 Years from the Infant Health and Development Program. *Jama 272*(16), 1257–1262.

Brown, C.

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

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

Burke, L. M. and D. B. Muhlhausen (2013). Head Start Impact Evaluation Report Finally Released. *The Heritage Foundation (ed.), Issue Brief 3823*.

Cahan, E. D. (1989). *Past caring: A history of US preschool care and education for the poor, 1820–1965.* National Center for Children in Poverty.

Campbell, F., G. Conti, J. J. Heckman, S. H. Moon, and R. Pinto (2014a). Early childhood program improves adult health. *Science*.

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

Campbell, F., G. Conti, J. J. Heckman, S. H. Moon, R. Pinto, E. Pungello, and Y. Pan (2014a). Early Childhood Investments Substantially Boost Adult Health. *Science 343*(6178), 1478–1485.

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

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

Campbell, F. A. and C. T. Ramey (1994). Effects of early intervention on intellectual and academic achievement: a follow-up study of children from low-income families. *Child development 65*(2), 684–698.

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

Campbell, F. A., B. H. Wasik, E. Pungello, M. Burchinal, O. Barbarin, K. Kainz, J. J. Sparling, and C. T. Ramey (2008). Young Adult Outcomes of the Abecedarian and CARE Early Childhood Educational Interventions. *Early Childhood Research Quarterly 23*(4), 452–466.

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. Technical report, National Bureau of Economic Research.

Caucutt, E. M. and L. J. Lochner (2011). Early and late human capital investments, borrowing constraints, and the family. Unpublished manuscript, University of Western Ontario. Retrieved from http://economics.uwo.ca/faculty/lochner/papers/earlyandlatehumancapital_Oct12.pdf.

Caucutt, E. M. and L. J. Lochner (2012). Early and late human capital investments, borrowing constraints, and the family. Working Paper.

Chafe, W. H. (1981). *Civilities and Civil Rights: Greensboro, North Carolina, and the Black Struggle for Freedom*. Oxford University Press.

Condon, S. (2015). Obama: We had universal childcare in the 1940's, so let's do it in 2015.

Cunha, F. (2007). *An Essay on Skill Formation*. Ph. D. thesis, University of Chicago.

Cunha, F. (2012). Eliciting maternal beliefs about the technology of skill formation. Presented at, Family Inequality Network: Family Economics and Human Capital in the Family, November 16, 2012.

Cunha, F., I. Elo, and J. Culhane (2013). Eliciting maternal expectations about the technology of cognitive skill formation. Technical report, 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., J. J. Heckman, L. J. Lochner, and D. V. Masterov (2006). Interpreting the evidence on life cycle skill formation. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Chapter 12, pp. 697–812. Amsterdam: North-Holland.

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). Early Childhood Education Programs. *Journal of Economic Perspectives*, 213–238.

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

Currie, J. and D. Thomas (1999). Does Head Start Help Hispanic Children? *Journal of Public Economics 74*(2), 235–262.

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

Del Boca, D., C. Flinn, and M. Wiswall (2013). Household choices and child development. Forthcoming.

Del Boca, D., C. Flinn, and M. Wiswall (2014a). Household choices and child development. *The Review of Economic Studies 81*(1), 137–185.

Del Boca, D., C. Flinn, and M. Wiswall (2014b). Household choices and child development. *The Review of Economic Studies 81*(1), 137–185.

Deming, D. (2009). Early Childhood Intervention and Life-cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics 1*(3), 111–134.

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

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.

Ellwood, D. T. and C. Jencks (2004). The spread of single-parent families in the united states since 1960.

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.

Fitzpatrick, M. D. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The BE Journal of Economic Analysis & Policy 8*(1).

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

García, J. L. and J. J. Heckman (2014). Ability, Character, and Social Mobility.

García, J. L., A. Hojman, and J. K. C. Shea (2014). Parental Time Allocation and Children's Cognitive and Non-cognitive Development.

Gayle, G.-L., L. Golan, and M. A. Soytas (2013). What accounts for the racial gap in time allocation and intergenerational transmission of human capital.

Gensowski, M., M. Griffin, and J. J. Heckman (2014). Effectiveness of chicago child parent centers in producing young-adult outcomes. *Unpublished Manuscript*.

Gormley, 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).

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.

Gray, S. W. (1969). Selected longitudinal studies of compensatory education–a look from the inside.

Gray, S. W. and R. A. Klaus (1965). An experimental preschool program for culturally deprived children. *Child Development*, 887–898.

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

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

Hart, B. and T. R. Risley (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: P.H. Brookes.

Haskins, R. (2010, October 11). Finally, Obama Administration is Putting Head Start to the Test.

Havnes, T. and M. Mogstad (2011a). Money for nothing? universal child care and maternal employment. *Journal of Public Economics 95*(11), 1455–1465.

Havnes, T. and M. Mogstad (2011b). No child left behind: Subsidized child care and children's long-run outcomes. *American Economic Journal: Economic Policy*, 97–129.

Havnes, T. and M. Mogstad (2014). Is universal child care leveling the playing field? *Journal of Public Economics*.

Heckman, J., S. H. Moon, R. Pinto, P. Savelyev, and A. Yavitz (2010a). Analyzing Social Experiments as Implemented: a Reexamination of the Evidence from the HighScope Perry Preschool Program. *Quantitative Economics 1*(1), 1–46.

Heckman, J. J. (1974). Effect of Childcare Programs on Women's Work Effort. *Journal of Political Economy*, S136–S163.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 153–161.

Heckman, J. J. (2008). Schools, Skills, and Synapses. *Economic Inquiry 46*(3), 289–324.

Heckman, J. J. (2012). *Giving Kids a Fair Chance (A Strategy That Works)*. Cambridge, MA: MIT P.

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. and T. Kautz (2012). Hard evidence on soft skills. *Labour economics 19*(4), 451–464.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3A, Chapter 31, pp. 1865–2097. New York: North-Holland.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010b). The Rate of Return to the HighScope Perry Preschool Program. *Journal of Public Economics 94*(1), 114–128.

Heckman, J. J. and S. Mosso (2014). The economics of human development and social mobility. *Annual Review of Economics 6*(1), 689–733.

Heckman, J. J., R. Pinto, and P. Savelyev (2013). Understanding the Mechanism through which and Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review 103*(6), 2052–86.

Heckman, J. J., R. Pinto, A. Shaikh, and A. Yavitz (2010). Inference with imperfect randomization: The case of the Perry Preschool Program. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and B. Setzler (2014). A Reanalysis of the Infant Health and Development Program. Forthcoming.

Heckman, J. J. and S. Urzua (2010). Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics 156*(1), 27–37.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144. Amsterdam: Elsevier.

Hill, C., W. Gormley, and S. Adelstein (2012). Do the Short-term Effects of a Strong Preschool Program Persist. *Center for Research on Children in the United States Working Paper 18*.

Hojman, A. (2015). Evidence on the fade-out of iq gains from early childhood interventions: a skill formation perspective. Working paper, University of Chicago.

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

Klein, J. (2011, July 7). Time to Ax Public Programs That Don't Yield Results.

Kline, P. and C. Walters (2014). Evaluating Public Programs with Close Substitutes: The Case of Head Start.

Kottelenberg, M. J. and S. F. Lehrer (2013). New evidence on the impacts of access to and attending universal child-care in canada. *Canadian Public Policy 39*(2), 263–286.

Lareau, A. (2011). *Unequal Childhoods: Class, Race, and Family Life* (2 ed.). Berkeley, CA: University of California Press.

Lee, S. Y. T. and A. Seshadri (2014). On the intergenerational transmission of economic status. *Unpublished manuscript, University of Wisconsin–Madison, Department of Economics*.

Lipsey, M., K. Hofer, N. Dong, D. Farran, and C. Bilbrey (2013a). Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow? Up Results from the Randomized Control Design. Technical report, Peabody Research Institute.

Lipsey, M., K. Hofer, N. Dong, D. Farran, and C. Bilbrey (2013b). Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow? Up Results from the Randomized Control Design. Technical report, Peabody Research Institute.

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, et al. (2002). Making a Difference in the Lives of Infants and Toddlers and Their Families: the Impacts of Early Head Start. Vol. I: Final Technical Report. Princeton, NJ: Mathematica Policy Research. Technical report, Mathematica Policy Research.

Ludwig, J. and D. L. Miller (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The 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.

Masse, L. N. and W. S. Barnett (2002). *A Benefit Cost Analysis of the Abecedarian Early Childhood Intervention*. National Institute for Early Education Research New Brunswick, NJ.

McCarton, C. M., J. Brooks-Gunn, I. F. Wallace, C. R. Bauer, F. C. Bennett, J. C. Bernbaum, R. S. Broyles, P. H. Casey, M. C. McCormick, D. T. Scott, et al. (1997). Results at age 8 years of early intervention for low-birth-weight premature infants: The Infant Health and Development Program. *Jama 277*(2), 126–132.

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, et al. (2006). 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.

Muhlhausen, D. B. and D. Lips (2010). Head Start Earns an F: No Lasting Impact for Children by First Grade. *The Heritage Foundation*.

NIEER (2013). The state of preschool 2013.

Paulsell, D., J. Cohen, A. Stieglitz, E. Lurie-Hurvitz, E. Fenichel, and E. Kisker (2002). *Partnerships for quality: Improving infant-toddler child care for low-income families*.

Pinto, R. (2014). Selection bias in a controlled experiment: The case of moving to opportunity. Technical report, Mimeo., University of Chicago.

Puma, M., S. Bell, R. Cook, C. Heid, G. Shapiro, P. Broene, F. Jenkins, P. Fletcher, L. Quinn, J. Friedman, et al. (2010). Head Start Impact Study. Final Report. *Administration for Children & Families*.

Ramey, C., G. McGinness, L. Cross, A. Collier, S. Barrie-Blackley, and K. Borman (1979). The Abecedarian approach to social competence: Cognitive and linguistic intervention for disadvantaged preschoolers.

Ramey, C. T. et al. (1979). The Abecedarian Approach to Social Competence: Cognitive and Linguistic Intervention for Disadvantaged Preschoolers.

Ramey, C. T., D. M. Bryant, B. H. Wasik, J. J. Sparling, K. H. Fendt, and L. M. La Vange (1992). Infant Health and Development Program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics 89*(3), 454–465.

Restuccia, D. and C. Urrutia (2004, December). Intergenerational persistence of earnings: The role of early and college education. *American Economic Review 94*(5), 1354–1378.

Reynolds, A. J., J. A. Temple, B. A. White, S.-R. Ou, and D. L. Robertson (2011). Age 26 Cost-Benefit Analysis of the Child-Parent Center Early Education Program. *Child Development 82*(1), 379–404.

Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores (2005). *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*. Ypsilanti, MI: High/Scope Press.

Shonkoff, J. P. and S. J. Meisels (1990). Early childhood intervention: The evolution of a concept. *Handbook of early childhood intervention*, 3–31.

Steve Aos and Roxanne Lieb and Jim Mayfiel and Marna Miller and Annie Pennucci (2004, September). Benefits and Costs of Prevention and Early Intervention Programs for Youth. Technical report, Washington State Institute for Public Policy.

The White House (2013)). President Obamas Plan for Early Education for all Americans. http://www.whitehouse.gov/thepressoffice. Accessed: 2014/09/16.

Torcasso, J. (2014). Head start program and the head start impact study. Working paper, University of Chicago.

U.S. Department of Education (2009). IDEA Programs. http://www2.ed.gov/programs/oseppsg/index.html. Accessed: 2014/09/19.

U.S. Department of Education (2012). ARRA Fact Sheet. http://www2.ed.gov/policy/gen/leg/recovery/factsheet/idea.html. Accessed: 2014/09/18.

Vogel, C. A., N. Aikens, A. Burwick, L. Hawkinson, A. Richardson, L. Mendenko, and R. ChazanCohen (2006). Findings from the survey of early head start programs: Communities, programs, and families. final report. *US Department of Health and Human Services*.

Walters, C. (2014). Inputs in the production of early childhood human capital: Evidence from head start. Technical report, National Bureau of Economic Research.

Wasik, B. H., C. T. Ramey, D. M. Bryant, and J. J. Sparling (1990). A longitudinal study of two early intervention strategies: Project CARE. *Child development 61*(6), 1682–1696.

Weikart, D. P. (1967). Preliminary results from a longitudinal study of disadvantaged preschool children. Unpublished manuscript, ERIC No. ED 030 490. Presented at the 1967 convention of the Council for Exceptional Children, St. Louis, MO.

Weikart, D. P. (1970). Longitudinal Results of the Ypsilanti Perry Preschool Project. Final Report. Volume II of 2 Volumes. Technical report.

Weikart, D. P., J. T. Bond, and J. T. McNeil (1978). *The Ypsilanti Perry Preschool Project: Preschool years and longitudinal results through fourth grade*. High/Scope Educational Research Foundation.

Weikart, D. P., C. K. Kamii, and N. L. Radin (1964). Perry Preschool Project: Progress Report. Technical report, Ypsilanti Public Schools, Ypsilanti, Mich.

Zigler, E. and S. Muenchow (1994). *Head Start: The inside story of America's most successful educational experiment*. Basic Books.

Zigler, E. and S. J. Styfco (2010). *The Hidden History of Head Start*. Oxford University Press.

A List of Abbreviations

ABC	Carolina Abecedarian Project
BA	Bachelors Degree
CARE	Carolina Approach to Responsive Education
CAT	California Achievement Test
CBA	Cost Benefit Analysis
CBCL	Achenbach Child Behavior Checklist
CBI	Classroom Behavior Inventory
CCDF	Child Care Development Fund
CCDP	Comprehensive Child Development Program
CD	Cultural Deprivation Scale
cNLSY79	Children of the National Longitudinal Survey of Youth 1979
CPC	Chicago Child-Parent Center
EHS	Early Head Start
EEP	Early Education Program
ETP	Early Training Project
HLBW	Higher Low Birth Weight (IHDP)
HS	Head Start
HSIS	Head Start Impact Study
IBR	Bayley's Infant Behavior Record
IHDP	Infant Health and Development Program
IQ	Intelligence Quotient
ISS	Intensive Substudy
ITT	Intent to Treat
K&R	Kohn and Rosman Test Behavior Instrument
LLBW	Lighter Low Birth Weight
LATE	Local Average Treatment Effect
LPN	Licensed Practical Nurse
MA	Masters Degree
MAT	Metropolitan Achievement Test
MSW	Master of Social Work
NCAST	Parent-Child Interaction Scales
NIEER	National Institute for Early Education Research
NLSY79	National Longitudinal Survey of Youth 1979
NSO	National Study Office (for IHDP)
PBI	Pupil Behavior Inventory
PDO	Program Development Office (for IHDP)
PFA	Preschool for All
PPP	Perry Preschool Program
PPVT	Peabody Picture Vocabulary Test
PPVT-R	Peabody Picture Vocabulary Test Revised
PSID	Panel Study of Income Dynamics
RAPS	Research Assessment Package for Schools
RCT	Randomized Controlled Trial
RN	Registered Nurse
SB	Stanford-Binet Intelligence Scale
SD	Standard Deviation
TN-VPK	Tennessee Voluntary Pre-K Program
TVIP	Test de Vocabulario en Imagenes Peabody (PPVT test for Spanish-speaking and bilingual students)
WIS	Wechsler Intelligence Test
WISC	Wechsler Intelligence Scale for Children
WJ	Woodcock Johnson Test of Achievement
YRS	Ypsilanti Rating Scale

B Methodology

This appendix presents a methodology enabling us to explore the treatment effects of the demonstration programs we analyze on early-life cognitive and character skills and later-life economic outcomes.

B.1 Model

Let D_i denote a treatment indicator for individual i, where i is a generic individual in the set $\mathcal{I}$. Let M_i be a male indicator and X_i a vector of background characteristics. D_i takes the value $d \in \{0, 1\}$ for individuals in the control or treatment groups, respectively. Let $Y_{i,s}^d$ denote outcome s for individual i where $\mathcal{S}$ denotes the set of outcomes and $s \in \mathcal{S}$ is a generic outcome, for $d \in \{0, 1\}$. Thus, the counterfactual outcome s for individual i is

$$Y_{i,s} = Y_{i,s}^1 D_i + Y_{i,s}^0 (1 - D_i) . \tag{12}$$

For $s \in \mathcal{S}$, our objective is to estimate the coefficients in the following linear model

$$Y_{i,s} = \beta_0^s + \beta_M^s M_i + \beta_D^s D_i + \beta_{MD}^s M_i D_i + \varepsilon_{is} \tag{13}$$

where ε_{is} is an error term. We suppress X_i to simplify notation but we control for it in all of our empirical analyses. As explained in Section B.2, conditioning on certain covariates is especially important in the analysis of ABC and PPP.

Table B.1: Parameters of Interest

Treatment Effect, Male	Treatment Effect, Female	Treatment Effect, Gender Difference: Male - Female
β_D^s	$\beta_D^s + \beta_{MD}^s$	β_{MD}^s

Note: this table lists the meaning of the parameters in (13) for each outcome. We estimate (13) for a collection of parameters $\mathcal{S}$, with typical element s.

Thus, our empirical objective is to test whether β_D^s is statistically and economically significant for females and likewise $\beta_D^s + \beta_{MD}^s$ for males. We also present estimates pooling females and

males. This is useful in this context given the small sample sizes. However, some results may be sex specific.

B.2 Estimation Issues

We face four issues when estimating and performing inference on the coefficients of (13): (i) compromised randomization; (ii) small sample size; (iii) multiple hypothesis testing; (iv) item non-response.

The first arises in ABC and PPP from the initial randomization: some children were swapped between treatment and control due to their background characteristics as explained in Section 4. The second is due to the design of the program. Thus, usual asymptotic inference methods generate imprecise estimations for the p-value of each test.[21]

The third issue arises from the nature of our research question. We will test the same hypothesis for various outcomes that seem to capture a lot of the impact of these programs. Thus, some researchers would be curious for us to correct the inference for multiple hypothesis testing. Roughly speaking, we need to correct for the fact that the probability of incurring in Type I error increases as the number of hypotheses to test increase. We think of the hypotheses we test as independent to each other, so we also present uncorrected p-values.

Finally, there are two ways in which we could observe item non-response. Either we do have data on follow up at the relevant outcome age but we miss information on a particular question or there is attrition in the sense that we do not observe data at all. Methodologically, we treat both of these as item non-response.

To solve these four issues, we use the methodology in Campbell et al. (2014a). Roughly speaking, we do the following. To solve the first issue, we control for the background characteristics altering the randomization design detailed in Section 4. In other words, we work under a "matching on observed variables" assumption. To solve the small sample size issue, we use a bootstrap based inference to avoid relying on asymptotic inference methods. In

[21] Heckman et al. (2010) develop a more robust procedure.

brief, our inference method is non-parametric. To solve the third issue, we correct multiple hypothesis testing through a "step-down procedure". Lastly, to solve the non-response issue, we use an inverse probability scheme.

In all of our estimations, we report two estimates of the p-value. The first corrects for all the estimation issues except for multiple hypothesis testing. The second corrects for all of the estimation issues. We label the first as "non-parametric p-value" and the second "non-parametric, step-down p-value". Table B.2 makes this explicit.

Table B.2: What Estimation Issues Does Each Type of p-value Address?

$p-value$	Compromised Randomization	Small Sample	Multiple Hypotheses Testing	Item Non-response
Non-parametric	✓	✓	×	✓
Non-parametric, Step-down	✓	✓	✓	✓

Note: this table details the estimation issues each of the two p-value's we present addresses.

C Data Sources

In order to explore the treatment effects on early-life outcomes of the demonstration programs, we generate factors to act as measures of several outcomes of interest. In particular, we factor analyze cognitive and character outcomes during the early ages. We choose age 8 to be our point of comparison because this is the age at which our data is most comparable across our five programs: PPP, ABC, CARE, IHDP, and ETP. Tables C.1 and C.2 list the instruments we use to generate the early age outcomes.

As our measure of cognitive ability, we use standardized IQ and achievement tests scores. In the case of PPP and ETP, achievement tests scores are converted from their raw form to their standardized forms at the University of Chicago Test Center.

To measure the non-cognitive early childhood outcomes, we follow the Big Five paradigm and construct factors measuring openness to experience, conscientiousness, extroversion, agreeableness, and emotional stability.

- **Openness to experience** refers to the child's tendency to be receptive of new aesthetic, cultural or intellectual experiences.
- **Conscientiousness** reflects the child's tendency to be organized, responsible, and hardworking.
- **Extroversion** is a measure of the child's preference for the outer world of people over the inner world of subjective experience.
- **Agreeableness** is a measure of the child's cooperation and altruism.
- **Emotional stability** is most easily understood as the reversal of neuroticism, which is an assessment of how prone an individual is to psychological distress. Neurotic individuals are easily stressed or upset, whereas emotionally stable ones are more resolute and resilient in their emotions.

To generate the factors measuring each trait described above, question items from instruments measuring personality traits are analyzed and designated to a single trait. Each group of items is then standardized and factor analyzed using the principal factor method, retaining only one factor. Mean imputation is applied to items with missing values after standardization, and prior to factor analysis. For each program, this procedure is repeated for every age that a personality test was administered. Similar to IQ and achievement, we then factor analyze the factors measuring each trait across the ages available, generating another set of character factors that is representative of the early years of the children's lives.

Table C.1: Sources for Outcomes at age 8 in ABC and CARE

Measure	ABC	CARE
IQ	WIS	WISC
Achievement	CAT, PIAT, WJ	WJ
Openness	CBI, CBCL Teacher, Walker	CBI, CBCL Teacher
Conscientiousness	CBI, CBCL Teacher, Walker	CBI, CBCL Teacher
Extroversion	CBI, CBCL Teacher, Walker	CBI, CBCL Teacher
Agreeableness	K&R, CBI, CBCL Teacher, Walker	K&R, CBI, CBCL Teacher
Emotional Stability	K&R, CBI, CBCL Teacher, Walker	K&R, CBI, CBCL Teacher

Note: **IQ Tests:** SB, Stanford Binet Intelligence Scale; WIS, Weschler Intelligence Scale; WISC, Wechsler Intelligence Scale for Children. **Achievement Tests:** CAT, California Achievement Test; MAT, Metropolitan Achievement Test; PIAT, Peabody Individual Achievement Test; WJ, Woodcock-Johnson Test of Achievement. **Personality Tests:** CBCL, Achenbach Child Behavior Checklist; CBI Classroom Behavior Inventory; IBR, Bayley's Infant Behavior Record; K&R, Kohn and Rosman Test Behavior Instrument; PBI, Pupil Behavior Inventory; RAPS, Research Assessment Package for Schools; Richman, Richman Child Behavior Problems Checklist; Walker, Walker Problem Behavior Identification Checklist; YRS, Ypsilanti Rating Scale.

Table C.2: Sources for Outcomes at age 8 in ETC, IHDP, and PPP

Measure	ETP	IHDP	Perry
IQ	SB	WISC	SB
Achievement	MAT	WJ	CAT
Openness	Teaching Rating	Richman, CBCL Parent, Teacher Survey, RAPS Teacher	PBI, YRS
Conscientiousness	Teaching Rating, Reputation Test	IBR, CBCL Parent, Teacher Survey, RAPS Teacher	PBI
Extroversion	Teaching Rating, Reputation Test	IBR, CBCL Parent, Teacher Survey	PBI YRS
Agreeableness	Teaching Rating, Reputation Test	Richman, CBCL Parent, Teacher Survey, RAPS Teacher	PBI
Emotional Stability	Teaching Rating, Reputation Test	Richman, CBCL Parent, Teacher Survey, RAPS Teacher	PBI, YRS

Note: **IQ Tests:** SB, Stanford Binet Intelligence Scale; WIS, Weschler Intelligence Scale; WISC, Wechsler Intelligence Scale for Children. **Achievement Tests:** CAT, California Achievement Test; MAT, Metropolitan Achievement Test; PIAT, Peabody Individual Achievement Test; WJ, Woodcock-Johnson Test of Achievement. **Personality Tests:** CBCL, Achenbach Child Behavior Checklist; CBI Classroom Behavior Inventory; IBR, Bayley's Infant Behavior Record; K&R, Kohn and Rosman Test Behavior Instrument; PBI, Pupil Behavior Inventory; RAPS, Research Assessment Package for Schools; Richman, Richman Child Behavior Problems Checklist; Walker, Walker Problem Behavior Identification Checklist; YRS, Ypsilanti Rating Scale.