Dougan, William; García, Jorge Luis; Polovnikov, Illia

Working Paper

High-Quality Early-Childhood Education at Scale: Evidence from a Multisite Randomized Trial

IZA Discussion Papers, No. 16442

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Dougan, William; García, Jorge Luis; Polovnikov, Illia (2023) : High-Quality Early-Childhood Education at Scale: Evidence from a Multisite Randomized Trial, IZA Discussion Papers, No. 16442, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/279140

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16442

High-Quality Early-Childhood Education at Scale: Evidence from a Multisite Randomized Trial

William Dougan
Jorge Luis García
Illia Polovnikov

SEPTEMBER 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16442

High-Quality Early-Childhood Education at Scale: Evidence from a Multisite Randomized Trial

William Dougan
Clemson University

Jorge Luis García
Clemson University and IZA

Illia Polovnikov
Boston College

SEPTEMBER 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

High-Quality Early-Childhood Education at Scale: Evidence from a Multisite Randomized Trial

We offer a new analysis of a large-scale trial of an early-childhood education program that targeted premature, low-birthweight children. This targeting heavily oversampled twins, whose outcomes differed significantly from singletons'. Singletons' gains in short-term cognition and age-18 non-cognitive skills were comparable to those of the Perry Preschool and Carolina Abecedarian Projects, supporting those programs' scalability. For twins, however, the program generated smaller positive short-term gains and negative age-18 impacts. These outcome differences arise from differences in parents' response to the program. A household production model suggests that the possibility of jointly supplying parenting to twins helps explain those differences.

JEL Classification: J13, I28, C93, H43

Keywords: childcare, early childhood education, large-scale randomized trial, parental investment, parenting

Corresponding author:
Jorge Luis García
Clemson University
309-C Wilbur O. and Ann Powers Hall
Clemson, S.C. 29634
USA

E-mail: jlgarci@clemson.edu

1. Introduction

Analyses of the Perry Preschool and Carolina Abecedarian (ABC) Projects are essential in the justification of recent proposals by the federal government to increase the provision and quality of childcare and preschool. President Obama referred to the rate of return of the Perry Preschool Project in a State of the Union Address to promote his Preschool for All initiative (The White House, 2013b), which recent proposals by President Biden aim to solidify.[1] Public and academic debates question the policy relevance of research based on Perry and ABC. Whitehurst (2013) writes that "generalizations […] from research findings on Perry and Abecedarian are prodigious leaps of faith." Other critiques note that the programs' age and small sample sizes are major caveats preventing generalization. While Preschool for All is the source for these criticisms, new proposals have reinvigorated the debate (e.g., The Wall Street Journal Editorial Board, 2022).

At scientific stake is whether Perry and ABC, two demonstration programs with small-sample implementations, can be scaled up. While recent studies find that large-scale preschool programs can be effective (e.g., Bailey et al., 2021; Gray-Lobe et al., 2023), an analysis of a large-scale replication of Perry and ABC is relevant due to the importance of these two programs in academic and public debates. It is also relevant to the discussion of whether public policies preserve "voltage" when scaled (List, 2022). An attempt to replicate Perry and ABC at scale is the Infant Health and Development Program (IHDP), a multisite randomized trial implemented in eight states across the US. Current evidence on the effectiveness of IHDP is weak and it has been used to bolster the argument that Perry and ABC cannot be scaled effectively. When discussing IHDP, Murray (2013) concludes: "none of those first-rate programs [i.e., Perry and ABC] are replicable on a large scale."

The IHDP was a randomized trial targeting children born in 1985 who were disadvantaged as indicated by premature birth ($\leq$ 37 weeks) and low birthweight ($\leq$ 2,500 grams). It was of high quality by several standards. It treated the child participants and their parents from childbirth to age 3 in weekly home visits during the first year of intervention. It then treated child participants in intensive center-based childcare during the second and third years of intervention, in which home visits transitioned to a biweekly schedule. The IHDP was implemented in eight sites located in eight states of the US, forming samples of 100 to 137 children per site. Roughly one-third of these children were randomized into the treat-

[1] Recent proposals are Preschool for All (The White House, 2013a), Build Back Better (The White House, 2020), the American Families Plan (The White House, 2021). The justification of their early-education components is primarily based on Perry and ABC, referring to Heckman et al. (2010), García et al. (2020), and García et al. (2021).

ment group, receiving both of the program's components: home visits and childcare. The remaining children formed the control group.

The eligibility criteria of IHDP produced a *de facto* oversampling of twins, who are more likely to be premature and of low birthweight than singletons (National Center for Health Statistics, 2022). Section 2 provides details of the IHDP and compares its programmatic content to Perry and ABC. Section 3 describes the data available for analysis. The data include primary data on the child participants of IHDP collected between birth and age 18. The data also include primary data on the child participants of Perry, which coincide in the age of observation with the IHDP data, allowing us to make treatment-effect comparisons between the two programs. We also compare the sample of the IHDP to the cohort of children born in 1985. In that cohort, 2% of children are twins, a much lower rate than the 10% observed in the IHDP sample. The comparison shows that the two eligibility criteria implemented in IHDP generated a sample of children who were at a socioeconomic disadvantage. For example, at baseline, 37% of the households of child participants participated in social programs, in contrast to the 3% observed for all households with children born in 1985.

We first evaluate the short-term effectiveness of the program, focusing on cognition at the end of the program. While impacts on cognition often fade out 2 or 3 years after programs end (Hojman, 2016), all successful early childhood education programs have a sizable effect on end-of-program cognition (Elango et al., 2016). Comparing end-of-program impacts across programs establishes a benchmark. An essential difference between previous studies and ours is that we make twinning an integral part of our analysis. This is important because outcomes of singletons and twins differ economically and statistically. For singletons, the end-of-program average treatment effect of IHDP on cognition is 9.3 points (no-effect p-value ≤ 0.05). We cannot reject the null hypothesis that this effect equals the end-of-program average treatment effect of 10.2 points generated by Perry (no-effect p-value ≤ 0.05)—both impacts are based on tests anchored to a national mean and standard deviation of 100 and 15 points. For twins, the end-of-program average treatment effect of IHDP is 4.3 points (no-effect p-value > 0.05). At a significance level of 10%, we reject the null hypothesis that the IHDP end-of-program effect for twins equals the effect of Perry. The oversampling of twins obscures the short-term effectiveness of the program.

Oversampling of twins also obscures the effectiveness of IHDP at boosting longer-term non-cognitive skills, as indicated by age-18 outcomes such as taking the SAT or ACT, not requiring special education, not being a smoker, or not being idle. These longer-term impacts speak to the life-cycle benefits of the program: they approximate academic motivation and externalizing behavior, which are the building blocks of treatment effects on life-cycle

education, labor income, and crime of programs like Perry and ABC (Conti et al., 2016; Heckman et al., 2013). Previous analyses finding weak longer-term impact of IHDP, used as evidence against the scalability of Perry and ABC, do not consider twinning. In contrast with previous studies, we find that for singletons the IHDP has an average treatment effect of up to 0.3 of a standard deviation on indices of the age-18 outcomes (no-effect p-value ≤ 0.05). For twins, the average treatment effect on these indices is negative. Section 4 elaborates on the analysis of treatment effects. It shows that estimates are robust to using different estimators and inferential procedures.

Section 5 begins our analysis of mechanisms. We focus on two inputs of the production function of children's skills, hours spent in childcare provided by the program (henceforth, childcare) and parental investment (henceforth, parenting), for which measures were obtained while the program was in place. Decomposition exercises show that these two inputs explain up to 65% of the short-term average treatment effect of the IHDP on cognition. We document that, upon treatment assignment, parents of singletons increase both childcare and parenting, while parents of twins increase childcare but decrease parenting. The parenting crowd-out for twins is substantial. Treatment increases our baseline measure of parenting by 0.2 of a standard deviation for singletons and it reduces it by 0.2 of a standard deviation for twins. To understand these differences in parental responses, in Section 6 we present a price-theoretic model of household production in which parents choose the level of welfare provided to their children and the mix of parenting and childcare used to produce that level.

Through that model, we seek an explanation of the observed differences in parental responses that does not depend on differences in preferences for child welfare or the cost of time. We allow for one essential difference in the production function of child welfare: parents of twins supply parenting to their twin children. Thus, at a given childcare-parenting combination, they are more productive than parents of singletons. This greater productivity leads to different optimal choices of input mixes and child-welfare levels by parents of twins and parents of singletons.

We treat the IHDP as a subsidy to childcare that generates three effects on the allocation of parents' time. Two of these effects are standard: a substitution effect toward greater child welfare relative to parents' own consumption and an income effect that increases both child welfare (and, thus, the inputs) and parents' own consumption. The third effect is a substitution effect in production towards childcare and away from parenting. We show that the ability of twins' parents to jointly provide parenting to both children can provide an explanation for different choices we observe that is not based on differences in the underlying production functions. We further argue that the sign and magnitude conditions required for

the income and substitution effects to generate the observed parental responses are plausible.

For parents of either singletons or twins, treatment assignment reduces parents' short-term cost of attaining any level of child welfare, so that an effective program would cause an increase in child welfare. Using age-3 cognition as the measure of child welfare, we find that the treatment effects of the IHDP were indeed positive, although those effects were smaller for twins than for singletons. This difference is explicable by the different substitution effects that a childcare subsidy entails for the two parental groups.

Once the program ends, the impact on twins becomes negative. We conjecture that the negative longer-term impact on twins may result from a rational response of parents who increase their labor force participation at the expense of their parenting. Using the additional resources generated by their additional work, they could monetarily compensate for the decrease in their children's welfare. It may also result from parents not being aware that the negative impact of the reduction of hands-on parenting is greater in the long run, once the program ends, than in the short run, when the program remediates the crowd-out. This latter explanation is consistent with studies documenting that parents, especially those at a socioeconomic disadvantage, underestimate the impact of their own parenting.

Section 7 finalizes our empirical analysis. We find that treatment increases our baseline measure of age-3 cognition by an average of 2/3 of a standard deviation for singletons, with 51% of that gain corresponding to childcare and 13% to parenting. For twins, childcare increases age-3 cognition by almost 2/3 of a standard deviation when holding parenting constant. The parenting crowd-out generated by treatment therefore substantially reduces the average treatment effect in magnitude and statistical significance. The decomposition relies on strong assumptions, which we challenge by using alternative methods for identifying the components that it requires. We cannot provide the decomposition for the age-18 outcomes because the smaller sample size generated by item non-response makes its estimation unreliable. Nonetheless, the consistent pattern of the treatment effects across outcomes suggests that, for twins, the crowd-out of parenting outweighs the positive impact of childcare on the age-18 outcomes. Section 8 summarizes and discusses policy implications.

2. The Infant Health and Development Program

The Infant Health and Development Program (Ramey et al., 1992) was designed to foster the development of children at socioeconomic disadvantage, measured by prematurity and low birthweight (Gross et al., 1997). The control and treatment groups of the IHDP received eight pediatric follow-ups when children were between 40 weeks and 36 months old. Additionally, the children in the treatment group received two main services: support to their parents

through home visits and high-quality center-based childcare.

2.1 Treatment Services

Home Visits. Professionals visited the households of the treated children, training the parents in problem-solving and parenting skills, following the curriculum *Partners for Learning*. Professionals demonstrated, practiced, and discussed the curriculum with parents to train them as partners in their children's learning. Parents were encouraged to reflect on daily life problems and to establish decision-making paths to solve such problems. Parents were prompted to observe, listen, and interact with their children. Home visits occurred weekly during the first year of treatment, transitioning to a bimonthly schedule in the second and third years of treatment.

High-Quality Childcare. Treatment at the childcare centers began when children reached age 1 and lasted two years. Children were able to spend between four and nine hours in childcare during weekdays, with the actual number of hours being chosen by their parents. The program organized transportation to and from each childcare center, which was utilized by more than 80% of the children. The childcare centers were for the exclusive use of the treated children and satisfied state licensing requirements. The teacher-child ratio was 1:3 during the first year of childcare treatment and decreased to 1:4 during the second year. Teachers were professionals, who not only continued *Partners for Learning*, but also introduced children to a set of activities based on the curriculum *Early Partners*, which aimed to foster sleep and awake states, eye-hand coordination, and independent handling and manipulation of objects, among other early life skills.

2.2 Comparison to Demonstration Programs

The IHDP was modeled after a pair of closely related programs trialed in the 1970s at the University of North Carolina at Chapel Hill: the Carolina Abecedarian Project (ABC; Ramey and Campbell, 1984) and the Carolina Approach to Responsive Education (CARE; Wasik et al., 1990). CARE was a continuation of ABC. Generally, ABC is analyzed separately, and CARE has received less attention due to its smaller sample. The programmatic elements of IHDP also align closely with those of the Perry Preschool Project. A loose interpretation of the IHDP is that it combined ABC and Perry. Like ABC, it started at birth. Like Perry, it included home visits and was not as intensive in its duration. Table 1 describes the details of the three programs, all of which had specific curriculum targets aiming to foster child development. Their curriculum and high adult-child ratios are essential in their classification as "high-quality."

Table 1. Details of the Infant Health and Development Program and Perry Preschool and Abecedarian Projects

	IHDP	Perry	ABC
Overview			
Years Implemented	1985-1988	1962-1968	1972-1985
Site	University Hospitals in 8 states	Ypsilanti, Michigan	Chapel Hill, North Carolina
Population Targeted	Low Birthweight or Prematurity	Disadvantaged African Americans	Disadvantaged, no race requirement
Cohorts	1	5	4
Age at Entry	0	3	0
Duration	3 years	2.5 years	5 years
Sample	377 treatment, 608 control	58 treatment, 65 control	58 treatment, 56 control (ABC)
Twins	42 treatment, 61 control	0	0
Main Treatment Components			
Home Visits	Once a week in first year Every other week after	4.5 per month	Not available
Center-based Care	All year, starting second year (20-45 hours per week)	30 weeks per year (12.5 hours per week)	50 week per year (40 hours per week)
Other Treatment Components	Basic health check-ups	None	Basic health check-ups and referrals Formula and diapers (also provided to controls) Nutrition
Substitutes Attended by Controls	Yes, after age 1	No	Yes, after age 2
Staff			
Adult-child Ratio	1:3 (1 to 2), 1:4 (2 to 3)	1:5 to 1:6	1:3 (0 to 1), 1:4.5 (1 to 4), 1:5.5 (age 4 to 5)
Teacher Certification	BA + 2 years of experience working with children under 3	BA	HS graduates mixes with certified staff
Other Specialists	Yes. Assistant teachers	No	Physician, nurse, social worker
Curriculum Targets			
Cognitive Development	Yes	Yes	Yes
High-risk Behavior	No	No	Yes
Language Development	No	Yes	Yes
Motor Development	No	No	Yes
Non-cognitive Development	Yes	Yes	Yes
School Readiness	No	Yes	Yes
Task Orientation	No	No	Yes
Health Status	Yes	No	Yes

Note: Details of the Infant Health and Development Program and Perry Preschool and Carolina Abecedarian Projects (IHDP, Perry, and ABC).
Sources: Authors' construction using Gross et al. (1993) for IHDP, Weikart et al. (1978) and Schweinhart et al. (1993) for Perry, and the appendix of García et al. (2020) for ABC.

2.3 Sample Formation and Characteristics

Initial Pool. An initial pool was formed with 4,551 mothers who gave birth to low-birthweight singleton children or twins between January and October of 1985 in eight university hospitals located across the United States.[2] Four factors disqualified 3,249 mothers from participating in the IHDP: (i) they were discharged before IHDP officers were able to contact them, (ii) their children were not premature, (iii) they gave birth to triplets or had higher-order multiple births, or (iv) they lived more than 45 minutes away from the university hospitals in which the childcare centers were located. The 1,302 remaining mothers were contacted by IHDP officers. Of these mothers, 1,028 agreed to participate in the randomization protocol once the program was described to them.

Randomization and Analysis Sample. The 1,028 mothers in the sample were stratified by state of residence and weight of their children—low-low ($\leq$ 2,000 grams) or low-high ($>$ 2,000 grams, $\leq$ 2,500 grams). They were then randomly assigned to either the treatment or control group, with a per-stratum probability of being treated equal to 1/3. Treatment status was assigned at the mother level. Twins were jointly assigned to either the treatment or control group. After randomization, 43 mothers withdrew from the program. It is not public information whether the withdrawers belonged to the treatment or control group (Gross et al., 1997). This source of attrition is minor, and we do not address it in this paper. The 985 remaining mothers gave birth to the children in the sample that we analyze.

Data are only available for one of the two siblings in all of the follow-ups for twin participants.[3] The sample of these followed-up twins together with the entire sample of singletons comprise our working analysis sample of children. We describe this sample at baseline in Table 2.[4] We observe 882 singletons (547 controls and 335 treatments) and 103 twins (61 controls and 42 treatments) at baseline. Originally, 880 children were classified as singletons and 105 children were classified as twins.[5] Panel *a* of Appendix Table A.1 displays the joint distribution of treatment and twinning status by state from baseline to age 3 for the analysis sample. Panels *b* to *e* display the distributions at the follow-ups. Attrition

[2]The hospitals or treatment sites were located in the states of Arkansas, Connecticut, Florida, Massachusetts, New York, Pennsylvania, Texas, and Washington.

[3]Data for both siblings in each twin pair are available from baseline to age three. For the later follow-ups, data on only one sibling in each twin pair are available. To keep the sample consistent throughout the paper, we only use the twins for whom data are available after age three. The twins for whom data are available after age three were picked randomly within each twin pair. The codebook reads: "For each [twin] pair only one twin (selected randomly with the result not known by caretaker or site staff) was a member [of the initial sample of 985 children]."

[4]Appendix Tables A.3 to A.5 provide summary statistics for the pooled sample and by sex at birth.

[5]Two cousin pairs were classified as twin pairs, but we classify them as singletons. One child within each cousin pair was followed up after age 3 (i.e., cousins were treated as twins for data-collection purposes).

decreases the sample sizes after age three. The estimators below address this attrition.

Sample Context. It is well documented that prematurity and low birthweight correlate with socioeconomic disadvantage.[6] Prematurity is the main cause of low birthweight and the primary cause of neonatal mortality in the US (National Center for Health Statistics, 2022). The IHDP's eligibility criteria led to oversampling of disadvantaged children relative to the population of children born in the US during 1985. The marriage rate at baseline of mother participants was 47%, compared to a rate of 78% among all women who gave birth in the US during 1985. The household-level take-up of social programs at baseline in the IHDP was 37%, while among all US households with newborn children it was 3%. Other measures of disadvantage display a similar pattern (see Table 2). Socioeconomic disadvantage and race also correlate in the US. As a consequence, the IHDP oversampled African-American children: 15% of newborns were African-American in 1985, while the percentage of African-American children in the IHDP was 52%.

The IHDP oversampled twins, who are up to eight times more likely to be premature and low-birthweight than singletons (National Center for Health Statistics, 2022). Unlike singletons, twins are likely to be premature and low-birthweight for biological reasons, holding all else equal.[7] The selection of mothers who accepted being part of the program is such that, within the sample of analysis at baseline, twinning displays a relatively low correlation with the characteristics listed in Table 2. Appendix Figure A.1 shows that the correlation of twinning with birthweight, gestational age, mother's education, and mother's age are at most 0.09 in absolute value in the treatment, control, and pooled samples. The figure also reports that a joint F-test of the relationship between twinning and all baseline characteristics in Table 2. The F-statistic is 1.04 (p-value = 0.40). Twins and singletons were thus at very similar socioeconomic disadvantage at baseline.

3. Data

3.1 Outcomes: Cognitive and Non-Cognitive Skill Measures

Cognitive. We observe measures of cognition from two different tests at ages 3, 5, 8, and 18 for the IHDP child participants. Panel a of Table 3 summarizes them. All of the tests observed are widely accepted and used as measures of cognitive skills. They have a

[6] See, for example, Almond et al. (2005), Aylward et al. (1989), Hoy et al. (1988), Leonard et al. (1990), Matte et al. (2001), McCormick et al. (1992), and Richards et al. (2001).

[7] Twinning has increased above the natural rate in the last thirty years, with the percent of newborns who are twins increasing from 2.3 in 1990 to 3.3 in 2019 (National Center for Health Statistics, 2022). This increase is due to the popularization of in-vitro fertilization methods, which in turn increase the probability of multiple births (Nassar et al., 2003). The IHDP's oversampling of twins was unlikely due to in-vitro fertilization methods, which were incipient in the early 1980s (Wang and Sauer, 2006).

Table 2. Baseline Characteristics of the Analysis Sample and Comparison to the Population of Children Born in 1985

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Infant Health and Development Program									
	Singletons				Twins					
	Control	Treatment	Difference	p -value	Control	Treatment	Difference	p -value	All	US in 1985
Panel a. **Children**										
Twin	0.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.10	0.02
Male	0.49	0.50	0.00	0.93	0.47	0.46	-0.01	0.91	0.49	0.51
Birthweight (grams)	1,792.10	1,817.52	25.42	0.44	1,749.08	1,804.78	55.70	0.47	1,798.50	3,350.44
Low Birth-weight	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	0.06
Gestational Age (weeks)	33.04	33.04	0.00	0.99	33.17	32.93	-0.24	0.60	33.04	41.62
African-American	0.52	0.51	-0.01	0.82	0.49	0.66	0.17	0.09	0.52	0.16
Hispanic	0.10	0.11	0.01	0.75	0.17	0.00	-0.17	0.00	0.11	0.16
Panel b. **Mother at Childbirth**										
African-American	0.52	0.51	-0.01	0.82	0.49	0.66	0.17	0.09	0.52	0.15
Age	24.90	24.30	-0.60	0.16	25.93	25.68	-0.25	0.82	24.80	25.82
Education	12.42	12.17	-0.26	0.14	12.78	12.68	-0.10	0.85	12.37	12.59
Works	0.36	0.35	-0.01	0.72	0.29	0.29	0.00	0.96	0.35	0.54
Married	0.49	0.42	-0.07	0.04	0.53	0.51	-0.01	0.90	0.47	0.78
Panel c. **Household at Childbirth**										
Use Social Programs	0.35	0.43	0.08	0.02	0.27	0.49	0.22	0.03	0.38	0.03
Siblings	0.78	0.71	-0.07	0.33	1.75	1.61	-0.14	0.51	0.85	0.85
Employed Adults	1.03	0.99	-0.04	0.46	0.93	1.15	0.21	0.13	1.02	1.18
Panel d. **Economy at Childbirth**										
Employment %	93.36	93.33	-0.03	0.78	93.27	93.17	-0.11	0.74	93.34	92.80
Median Income in 1000s (2020 USD)	55.52	55.51	-0.02	0.98	54.80	54.58	-0.23	0.90	55.43	54.33
Government Expenditure per Capita (2020 USD)	8,089.37	8,081.25	-8.12	0.93	8,005.94	8,001.94	-4.00	0.99	8,077.48	7,972.09
Panel e. **Joint Tests**										
F-statistic	0.82				3.33					
p-value	0.66				0.00					

Note: In Panels *a* to *d*, Columns (1) and (2) display the average baseline characteristics of the singletons in the sample by treatment status. Column (3) displays the difference between Columns (2) and (1). Column (4) displays the p-value of the t-statistic associated with the difference in Column (3). The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. Columns (5) to (8) are analogous in format to Columns (1) to (4) for the twins in the sample. Column (9) displays the average characteristics of the full sample, pooling the experimental groups and the singleton and twin children. Column (10) displays the average characteristics of children who were born during 1985 in the United States. Panel *e* presents the F-test corresponding to a joint test of significance in the mean-difference across all of the variables in Panels *a* to *d* for the samples of singletons and twins, as well as the pooled sample of singletons and twins. **Source:** For Column (10), the sources are Bureau of the Census (1986), Federal Reserve Bank of St. Louis (2022), IPUMS USA (2022), National Bureau of Economic Research (2022), Statista (2022), and US Bureau of Labor Statistics (2022).

straightforward interpretation because they are anchored to the mean and standard deviation of their national distribution, which are 100 and 15. To be clear, the in-sample average and standard deviation are not 100 and 15. Instead, a score of 110 on a test means that an individual is 2/3 of a standard deviation above the national mean of the test. This anchoring makes scores on tests across ages comparable.

The treatment and control averages are very similar within each age across the two observed tests. We therefore focus on the second test listed for each age. The second test is either the Stanford-Binet IQ test or a test similar in content. The tests observed allow for a precise comparison with Perry, for which we observe participants' Stanford Binet IQ test scores 0, 3, and 5 years after the ends of both programs.[8] Therefore, we observe test scores 0, 3, and 5 years after the end for both programs. For Perry, we do not observe test scores 15 years after the program ends. As we document below, treatment effects on test scores generally fade out or disappear 5 years after programs end. The comparison of cognitive outcomes 15 years after is thus less relevant than the comparison during prior years. At age 18, we focus on outcomes that approximate non-cognitive skills instead.

Non-Cognitive. Generally, the impact of early childhood education programs on cognition fades out a couple of years after those programs, including Perry and ABC, end (Hojman, 2016). That said, all successful early childhood education programs have a short-term (end-of-program) impact on cognition (Elango et al., 2016). Indeed, García and Heckman (2023) revisit the long-term impact of Perry and ABC on cognition and show that the apparent fade-out is a measurement artifact. While traditional IQ tests help measure short-term impacts, they do not help measure whether impacts persist. García and Heckman (2023) document that, when measured using modern methods, the impacts on cognition persist at least up to participants' age 54 (Perry) or age 45 (ABC).

We do not observe modern measures of cognition at age 18 for the IHDP participants. We observe outcomes approximating non-cognitive skills. We interpret "never in special education," "not having a reading tutor," "not having a math tutor," and "have taken the SAT or ACT" as proxies of academic motivation. We interpret "not being a smoker," "not being without a job or enrolled in school (idle)," "not being in therapy," and "not being a teen parent" as proxies of (lack of) externalizing behavior (i.e., behavior that could be problematic towards a person's environment). Heckman et al. (2013) show that the gains generated by Perry in terms of academic behavior and externalizing behavior explain its long-term impact

[8]Table 1 displays the sample details of Perry. A comparison to the impact of ABC would be less precise because tests are not observed during the same years after the program ends. Elango et al. (2016) show that the end-of-program impacts on cognition of Perry and ABC are very similar.

Table 3. Skill Measures by Treatment and Twinning Status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Singletons				Twins			
	Age	Observations	Control	Treatment	Difference	p-value	Control	Treatment	Difference	p-value
Panel a. **Cognitive**										
PPVT-R	3	812	85.11	91.39	6.28	0.00	85.24	86.50	1.26	0.73
Stanford-Binet IQ Test		908	81.47	90.81	9.34	0.00	81.29	85.64	4.35	0.26
PPVT-R	5	805	79.98	84.00	4.03	0.02	74.33	72.39	-1.94	0.69
Wechsler PPSI IQ Test		804	91.69	92.60	0.91	0.51	88.71	85.41	-3.30	0.35
PPVT-III	8	863	85.85	85.59	-0.26	0.88	82.07	79.38	-2.69	0.55
Wechsler ISC IQ Test		870	85.65	86.24	0.59	0.78	84.91	79.17	-5.74	0.30
PPVT-III	18	611	95.75	96.50	0.76	0.64	96.67	91.86	-4.82	0.31
Wechsler ASI IQ Test		614	91.53	91.81	0.28	0.85	91.87	89.93	-1.94	0.64
Panel b. **Non-Cognitive**										
Educational Outcomes										
Never in Special Education	18	888	0.70	0.74	0.04	0.24	0.67	0.69	0.02	0.80
No Reading Tutor		616	0.88	0.95	0.07	0.00	0.98	0.86	-0.12	0.09
No Math Tutor		618	0.76	0.81	0.04	0.22	0.86	0.75	-0.11	0.24
Took SAT or ACT		600	0.56	0.58	0.01	0.74	0.78	0.56	-0.23	0.05
Index: Average of Educational Outcomes		585	0.73	0.78	0.05	0.03	0.84	0.74	-0.09	0.11
Behavioral Outcomes										
Not a Smoker	18	605	0.75	0.79	0.04	0.28	0.85	0.81	-0.03	0.72
Not Idle		631	0.06	0.06	-0.00	0.96	0.02	0.07	0.05	0.36
Not in Therapy		622	0.69	0.71	0.02	0.66	0.76	0.54	-0.23	0.05
Not Teen Parent		598	0.86	0.85	-0.00	0.94	0.93	0.85	-0.08	0.29
Index: Average of Behavioral Outcomes		586	0.59	0.60	0.02	0.35	0.64	0.57	-0.08	0.07
All Outcomes										
Average of Indices	18	577	0.66	0.69	0.03	0.04	0.74	0.65	-0.09	0.03

Note: Column (1) displays the age of the participants when outcomes are measured. Column (2) reports the corresponding number of observations. Columns (3) and (4) display the average of the outcomes for the singletons in the sample by treatment status. Column (5) displays the difference between Columns (4) and (3). Column (6) displays the p-value of the t-statistic associated with the difference in Column (5). The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. Columns (7) to (10) are analogous in format to Columns (3) to (6) for the twins in the sample. The average of indices is the average of the two average or index outcomes.

on adulthood outcomes (e.g., criminal activity, employment, labor income).

Panel *b* of Table 3 describes the outcomes that approximate non-cognitive skills, all of which are indicator variables. To reduce dimensionality, we divide all outcomes into two categories ("educational" and "behavioral") and create corresponding indices by taking the average for each category. For each individual, the index is the average across the corresponding outcomes.[9] To construct the average of indices, we compute the raw average of the educational and behavioral indices for each individual.

Attrition and Non-Response. There was no attrition in either of the experimental groups from ages 0 to 3. The baseline characteristics and IQ test scores are fully observed at age 3 for 878 children using our least comprehensive control sets (Control Sets 1 and 2) and 872 children using our most comprehensive (Control Sets 3 and 4). The same is true for the childcare and parenting measures described below. The drop from the 985 children in the analysis sample is due to non-response. Without controls, the sample sizes are as indicated in Table 3. Except for "special education" (obtained from administrative records), the number of observations in the age-18 follow-up drops substantially because many individuals were not located for interviews. The estimators we use below address missing information due to attrition and item non-response (jointly referred to as attrition henceforth).

3.2 Inputs of the Production Function of Child Skills

Childcare. Mothers in the treatment and control groups reported the average of weekly hours that their children spent in childcare centers (we lump childcare centers and nurseries together and refer to them jointly as "childcare"). The mothers reported this average when their children were 18, 24, 30, and 36 months old, but did not report the specific center that their children attended. Qualitative descriptions from the IHDP's principal investigators and officers state that treatment-group children exclusively attended IHDP childcare centers, and their mothers remained committed and engaged with the program during the first 3 years of their children's lives (Gross et al., 1997). We therefore assume that the childcare hours that mothers reported were spent at the IHDP childcare centers. Although the control-group children were not in the IHDP childcare centers, their mothers were free to enroll them in alternative centers. We do not observe the quality of the centers attended by the children in

[9] Appendix Table A.2 shows that the treatment-control average differences and inference are similar when considering an alternative to these indices. This alternative aims to reinforce the quality of the indices as proxies of non-cognitive skills. We obtain the alternative indices by residualizing each of the indices from the two age-18 cognitive test scores in Table 3 using linear regressions. For instance, we average the four outcomes to construct the index or average of educational outcomes. We regress this average on the two age-18 IQ test scores. The residual of that regression is the relevant average or index we consider as the outcome and describe in Appendix Table A.2.

the control group, so our observations of the average weekly hours spent in childcare are not adjusted for quality.[10] We argue below that this lack of quality adjustment is only a minor concern in this context.

Parenting. The long form of the Home Observation Measurement of the Environment (HOME) inventory (Bradley and Caldwell, 1984; Bradley et al., 1992) was used to measure the interaction of the child and mother participants, as well as the resources available during that interaction. The HOME was measured when the child participants were 1 and 3 years old. The HOME inventories differ by age, to account for the natural process of child maturation. Panel *a1* of Table 4 describes the age-1 inventory, listing the inventory's six subscales. The score for each subscale is the average of several binary items that are exclusive to each subscale (i.e., each subscale has a dedicated measurement system of items). Linver et al. (2004) evaluate the correlation structure of each subscale's items and compare this structure across multiple datasets where the HOME subscale scores are observed. They find that each set of dedicated items yields a consistently similar correlation structure across diverse contexts and years.[11] Panel *b1* of Table 4 is analogous in format to Panel *a1* for the age-3 inventory.[12]

We use factor analysis to summarize the information in the conceptual subscales of the HOME inventories, enabling us to aggregate the subscales into a one-dimensional interpretable aggregate at each age. The analysis creates a latent factor variable that describes parenting for each mother-child pair, summarizes the covariability between the subscale scores, and accounts for measurement error by optimally combining the information in them. The scree tests implied by the eigenvalues in Panels *a2* and *b2* of Table 4 indicate that one latent factor variable appropriately represents all of the subscales at each age. The panels also display the loadings of each subscale in the latent factor variables, which indicate the importance of each subscale in the construction of the latent.[13] We estimate latent factor

[10]Mothers were asked for the primary and secondary care arrangements for their children. The answer options allowed them to report if they were primary or secondary caregivers, or if other individuals were (father or relatives). The options also allowed them to report if childcare centers or nurseries were the primary or secondary arrangements. If they reported using childcare centers or nurseries, they were asked questions allowing us to construct average weekly hours. The distinction between primary and secondary arrangements does not have a useful meaning. Mothers whose children spent most of the day in a childcare center could have reported this arrangement as "primary." Other mothers could have reported being the primary caregivers and, at the same time, could also have reported that their children spent most of the day in a childcare center, calling it a secondary arrangement. Our childcare variables lump together hours in childcare centers or nurseries, either if these were reported as primary or as secondary arrangements.

[11]The correlation structure of the items in each of the subscales is similar when items are measured in the Infant Health and Development Program, the NICHD Study of Early Childcare, the 1979 National Longitudinal Study of Youth-Child Supplement, and the Project of Human Development in Chicago Neighborhoods.

[12]Leventhal et al. (2004) provide a validation of the age-3 subscales analogous to Linver et al. (2004).

[13]The scree test is due to Cattell (1966). We construct it as follows: (i) Form as many factors as subscales

Table 4. Parental Investment or Parenting Measurement

Panel a. Age 1	Subscale:	Parental Warmth	Parental Verbal Skills	Parental Lack of Hostility	Learning/ Literacy	Activities/ Outgoings	Developmental Advance
Panel a1. **Brief Description**							
Number of Items		7	3	5	7	3	4
Example of an Item		Parent responds positively to praise of child offered by visitor	Parent converses freely and easily	Parent does not scold or criticize child during visit	Toys for literature and music	Child gets out of house at least four times/week	Parent structures child's play periods
Panel a2. **Factor Analysis**							
Factor	Eigenvalue	Loadings					
1	**1.66**	**0.61**	**0.46**	**0.16**	**0.69**	**0.32**	**0.68**
2	0.05	0.09	0.15	0.01	-0.09	-0.02	-0.08
3	0.02	-0.04	0.02	0.12	-0.00	-0.06	0.02

Panel b. Age 3	Subscale:	Learning Stimulation	Access to Reading	Parental Verbal Skills	Parental Warmth	Home Exterior	Home Interior	Outings/ Activities	Parental Lack of Hostility
Panel b1. **Brief Description**									
Number of Items		14	5	2	9	3	4	3	3
Example of an Item		Child has three or more puzzles	Child has access to at least 10 children's books	Parent uses complex sentence structure and vocabulary	Parent holds child close 10–15 min per day	Building appears safe	Rooms are not overcrowded with furniture	Child has been taken to a museum during the past year	Parent does not use physical restraint during visit
Panel b2. **Factor Analysis**									
Factor	Eigenvalue	Loadings							
1	**2.56**	**0.76**	**0.69**	**0.46**	**0.58**	**0.54**	**0.51**	**0.59**	**0.23**
2	0.35	-0.20	-0.20	0.03	-0.13	0.34	0.36	-0.04	0.02
3	0.08	-0.01	-0.07	0.06	0.14	-0.03	0.00	-0.14	0.18
4	0.02	0.02	-0.02	-0.10	-0.00	0.02	-0.01	0.04	0.06

Note: Panel *a* describes the measurement of parenting at age 1. Panel *a1* lists the subscales of the Home Observation Measurement of the Environment (HOME) inventory, measured at age 1. It then lists the number of (binary) items per subscale, as well as an example item. The subscales are scored as the average across items. Panel *a2* describes results from factor analyzing the subscales. We display the eigenvalues and subscale loadings of the first three factors. Panel *b* is analogous in format to Panel *a*. It describes the measurement of parenting at age 3.

variables for ages 1 and 3 and standardize them to an in-sample mean of 0 and a standard deviation of 1. These latents are our measures of parenting.

When verifying the assumption of the model in Section 6, we need to aggregate the inputs received by twin children within a household. For childcare, this operation is straightforward. Suppose the average number of hours a twin child attends childcare is 3.0. In that case, as captured in our data, the average number of hours a twin pair of children attends childcare is 6.0. For parenting, our factor measures cannot be aggregated, because they contain negative values (recall that they are standardized to an in-sample mean of 0 and standard deviation of 1). We thus create an alternative parenting measure by adding the binary items corresponding to the subscales of the HOME inventories at ages 1 and 3.[14] In either case, the parenting measure represents the parenting received by one child instead of a twin pair or the children in the household (thus, the need to aggregate within households with twins).

Interpretation of Input Measures. For the treatment-group children, the high quality of the childcare centers they attend is homogenous. Therefore, our hour variables are plausible measures of the resource content of childcare as an input to the production function of children's skills. For the control-group children, the quality of the childcare centers they attend, which we do not observe, may be heterogeneous. In this case, the resource content of our hour variables is uncertain. However, they spend relatively little time in childcare centers, even after including other types of childcare centers. Additionally, we mainly focus on treatment-control average differences, which are entirely driven by the take up of IHDP childcare. It is thus reasonable to assume that such a difference is primarily driven by high-quality resource content.

We also aim to measure the resource content of parenting. Measures of hours per week spent with mothers or others would fall short in measuring resource content if mothers or other caretakers were heterogeneous in quality. Using the HOME score aims to circumvent this issue, given that this tool produces a quality-adjusted measure of parenting by design. Some items of the HOME score measure material resources, while others measure one-on-one

and array them in a matrix; (ii) Calculate the eigenvalues of the matrix; (iii) Sort and label the factors increasingly according to their eigenvalues; (iv) Keep factors associated with an eigenvalue greater than one. The test is based on the intuition that factors associated with eigenvalues greater than one represent linearly independent combinations of the subscales. Panels *a1* and *b1* of Table 4 only show information on the factors with the three greatest eigenvalues. The remaining factors have much smaller eigenvalues. We omit them.

[14] Appendix Figure A.3 shows that the main treatment-effect results on parenting, displayed in Figure 5, are qualitatively identical when using this aggregate measure, rather than the factor variables. We use the factor variables throughout the paper, except in our discussion of Section 6, because measurement error is a concern when using HOME inventories.

mother-child interactions. Factor analyzing the sub-scales of the HOME scores produces a data-driven measure of the quantity and quality of parenting. Thus, we assume that our measure of parenting represents the resource content of a parenting unit. Studies estimating the production function of children's skills use the HOME scores similarly.[15,16]

One alternative to our approach is to measure the hours children spend with their mothers and other caretakers, like Chaparro et al. (2020), who use the IHDP and multiple other datasets to study the impact of several childcare policies. Appendix 2.1 discusses their parenting measures and provides a replication based on our data. Their measures are particularly useful when formally studying the parental time-allocation consequences of childcare policies, which is not the focus of our study. Of course, we may miss other inputs children receive during the day (e.g., care received by other caretakers, like fathers). However, childcare and parenting are likely the two main inputs children received during the ages we observed. Our analysis proceeds as if they were the only two inputs. The decomposition of treatment effects below indicates that these two inputs largely explain program impacts.

4. A Reevaluation of IHDP: Heterogeneity by Twinning Status

We analyze the impact of IHDP by twinning status. Let D_i be a binary indicator of the treatment status of child i, and denote that child's outcome when assigned to treatment status $d \in \{0(\text{control}), 1(\text{treatment})\}$ by Y_i^d. Using the switching regression of Quandt (1958, 1972), the child's observed outcome is

$$Y_i \quad = D_i \cdot Y_i^1 \quad + (1 - D_i) \cdot Y_i^0, \tag{1}$$

with average treatment effects

$$\text{ATE for Singletons} \quad := \quad \mathbb{E}\left[Y_i^1 - Y_i^0 | \text{child } i \text{ is a singleton}\right] \tag{2}$$

$$\text{ATE for Twins} \quad := \quad \mathbb{E}\left[Y_i^1 - Y_i^0 | \text{child } i \text{ is a twin}\right]. \tag{3}$$

[15]Several other studies implicitly assume that "all parenting" is contained in measures based on the HOME score (e.g., Agostinelli and Wiswall, 2016; Cunha and Heckman, 2008; Cunha et al., 2010; Todd and Wolpin, 2003, 2007). That does not mean that mothers or children do not have a time constraint. Instead, it means that the HOME summarizes their interactions as inputs of the production of child skills.

[16]Appendix Figures A.6 to A.8 show the treatment effect on the measures in Panels *c* and *d* of Table A.7. The program impacts on them correspond qualitatively with the impacts on parenting presented below. Appendix Figures A.9 and A.10 consider alternative measures of parenting that aim to further "quantity-adjust" our measure. One alternative is the residual from the regression of our main measure of parenting on maternal hours of care described in Appendix Table A.7. Another alternative residualizes not only maternal hours of care but also maternal working hours, aiming to account for all observed maternal time. The average treatment-control differences based on these alternatives remain essentially unchanged if compared to the difference based on our main measure.

Assignment to treatment is random. Further, twinning is a baseline characteristic. Therefore, within singletons and twins, assignment to treatment is also random. For brevity, we consider one estimator of the ATE in the main text (the raw or unadjusted average treatment-control difference or *mean difference*). We accompany the corresponding estimates with robust standard errors clustered at the child-participant level, as recommended by Abadie et al. (2023) for contexts like ours. Appendix Table A.8 shows that, across our main short-term and longer-term outcomes, adjusted mean differences obtained based on alternative estimators—ordinary least squares and inverse-probability weighting—yield estimates similar to the unadjusted mean differences.[17] It also shows that statistical tests yield the same conclusions when using alternative inferential methods.

4.1 Short-Term Impact

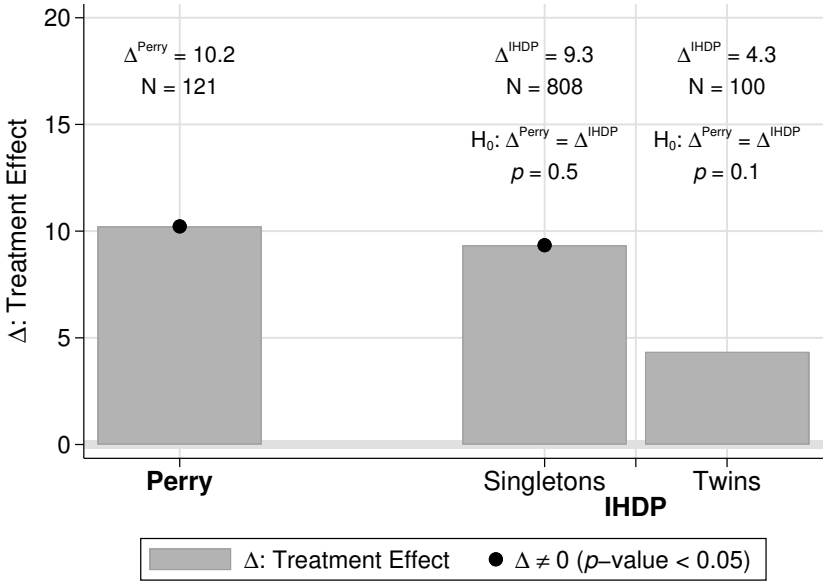
Figure 1a displays the end-of-program impact of IHDP on cognition by twinning status. For reference, it displays the same quantity for Perry. The impact of IHDP on singletons is about 2/3 of the national standard deviation in the test we use to measure cognition. For Perry, the impact is about the same. For IHDP twins, the impact is about half as much. We fail to reject the hypothesis that the impact of IHDP on singletons equals the impact of Perry. Using a 10% significance level, we reject the null hypothesis that the impact of IHDP on twins equals the impact of Perry. Additionally, the impact of IHDP on singletons differs statistically from 0 at a 5% significance level, while the impact of IHDP on twins does not. This latter impact does not differ statistically from 0 at a 10% significance level either.

Figure 1a is the first indication that pooling singletons and twins is misleading from a program-evaluation perspective: the majority of the IHDP sample is composed of singletons, and its impact on them is similar to the impact of Perry. When pooling singletons and twins, the overall impact is lower than Perry's impact. This could lead to the conclusion that, by scaling, IHDP achieves less than Perry by the end of the program (i.e., that scaling leads to a "voltage drop;" List, 2022).

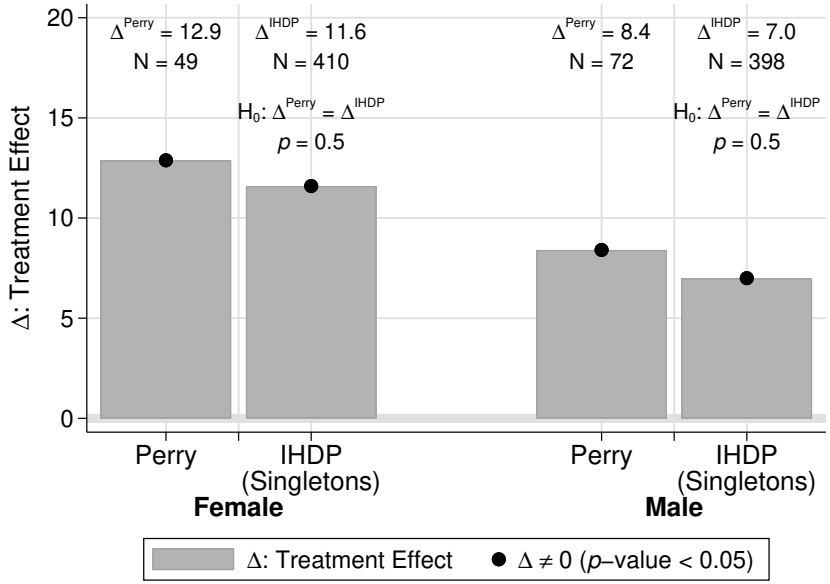
[17] When attrition is low, the mean difference is an appropriate estimator. However, there is substantial attrition for the age-18 outcomes (see Appendix Table 3). To address this issue, we use regression-adjusted mean differences estimated via ordinary least squares (OLS) as an alternative estimator. We do not know the exact reasons for attrition, but it imbalances the observed baseline characteristics between the treatment and control groups (see Appendix Table A.6). We adjust the average treatment-control difference by conditioning on the baseline variables in Panels *a*, *b*, and *c* of Table 2, as well as treatment site fixed effects. A second alternative estimator is the inverse-probability-weighting estimator (IPW; see Heckman and Karapakula, 2021), which weighs the regressions used to obtain the treatment-control mean difference by the inverse probability of being attrited. Weighting and imputation are based on the same baseline variables used for OLS. ATE estimates of Perry on short-term cognitive outcomes are based on the mean difference, an appropriate estimator given the low short-term levels of attrition and non-response.

Figure 1. End-of-Program Impact on Cognition: Perry and IHDP

(a) Perry and IHDP by Twinning Status



(b) Perry and IHDP Singletons by Sex



Note: Panel (a) displays the treatment effect (average treatment-control difference), denoted by Δ, for the pooled sample of participants of the Perry Preschool Program. It also displays the treatment effect for the singleton and twin participants of the Infant Health and Development Program (IHDP). For both programs, cognition is measured using the Stanford-Binet IQ test, anchored to its national mean of 100 and standard deviation of 15. For Perry, end-of-program is age 5. For IHDP, it is age 3. The plot labels the treatment effect and sample size. It also labels the treatment effect when the null hypothesis that the treatment effect is 0 is rejected using a significance level of 5%. Additionally, it displays the p-values for null hypotheses that the treatment effect of Perry equals the treatment effect of IHDP for singletons and twins. The p-values are based on robust standard errors clustered at the child-participant level. Panel (b) is analogous in format to Panel (a) by sex. For IHDP, this panel excludes twins.

The impact of Perry and ABC on end-of-program cognition, and, more generally, on longer-term skills and educational outcomes, is greater for female rather than male participants (Elango et al., 2016; García et al., 2018). Given the curricular similarities between these programs and IHDP, it is expected that, if IHDP is effective at scaling the success of these programs, the mechanisms driving differences by sex at birth prevail at scale. Figure 1b verifies this. Impacts are greater for female participants when compared to their male counterparts. By sex at birth, the impacts of Perry and IHDP are closely aligned.

An additional inquiry on IHDP's scaling is investigating its impact by site. Figure 2a displays such investigation for the IHDP singletons, relying on the end-of-program impact on cognition. While control and treatment levels vary across sites, the impacts have a tight range across seven of the eight sites, indicating that IHDP scaled the end-of-program effectiveness of Perry at boosting cognition. Despite small site-level samples, we reject the null hypothesis that the impact is 0 using a significance level of 5% in these seven sites. For twins, Figure 2b shows that the impacts are generally smaller and less precise across sites, consistent with the average impact for the full sample of them being smaller and non-significant.

Figures 3 and 2a indicate that, for the majority of its participants, IHDP was effectively scaled, as measured by end-of-program cognition. However, it is well-known that this impact fades out a couple of years after the program ends in the case of Perry. Figure 3a shows that the same is true for IHDP, even when focusing on singletons. Five years after the programs end, their impact on cognition is negligible. For Perry, Heckman et al. (2013) show that longer-term (age 40) program gains in outcomes such as education, labor income, and criminal activity are driven by gains in shorter-term skills other than cognition (i.e., externalizing behavior and academic motivation) measured between ages 7 and 9.

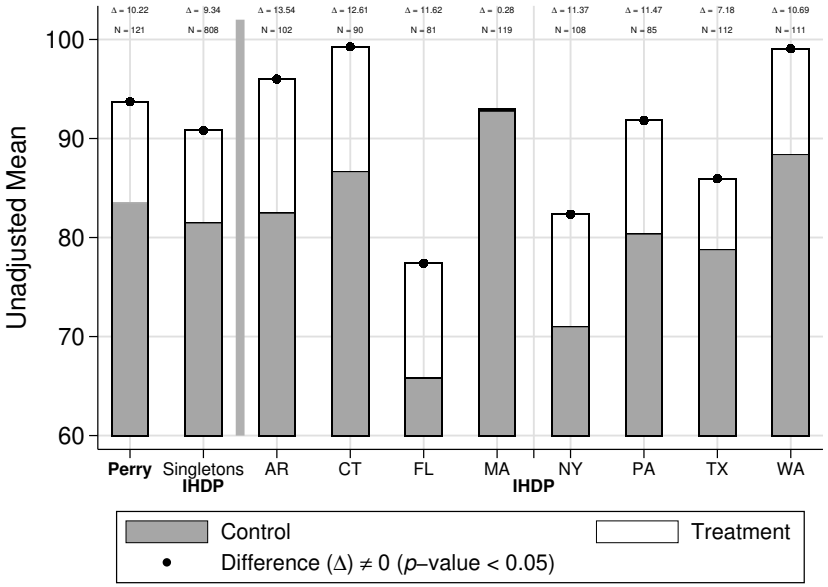
4.2 Longer-Term Impact on Non-Cognitive Skill Proxies

We do not observe skill measures during childhood for IHDP. However, we observe measures at age 18 resembling academic motivation and externalizing behavior.[18] These measures are the indices of educational and behavioral outcomes summarized in Section 3. We display the impact of IHDP on these outcomes for the pooled sample of singletons and twins in Figure 3b and barely find any impact. The results in Figure 3a lead to statements that IHDP is ineffective at scaling the success of programs like Perry and ABC (Murray, 2013). In addition, we speculate that the lukewarm results in Figure 3b have led to the scarce

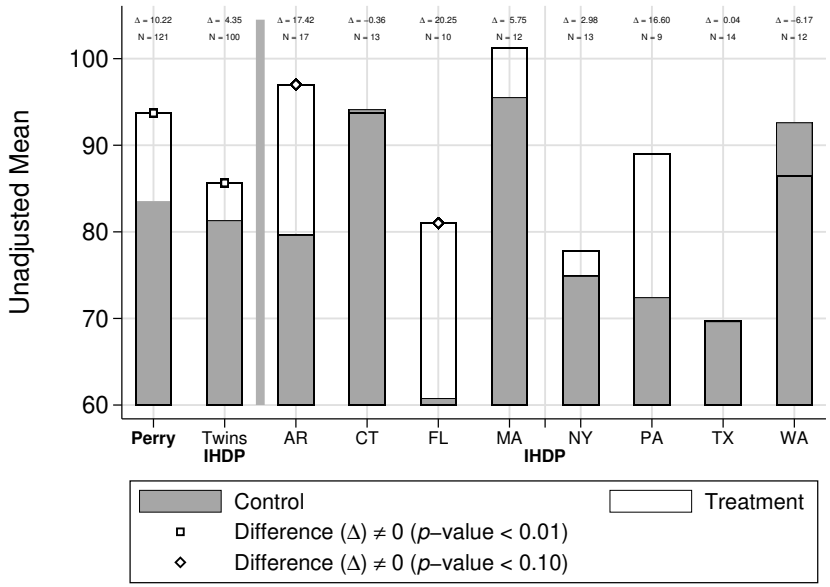
[18] Appendix Table A.6 shows that some average differences in baseline characteristics between those observed in the age-18 follow-up and those not observed are statistically significant. F-statistics associated with joint tests are also statistically significant, though small in magnitude. The similarity across estimates based on the different estimators of the ATE suggest that attrition is a relatively minor concern.

Figure 2. End-of-Program Impact on Cognition: Perry and IHDP by Site

(a) Perry and IHDP Singletons



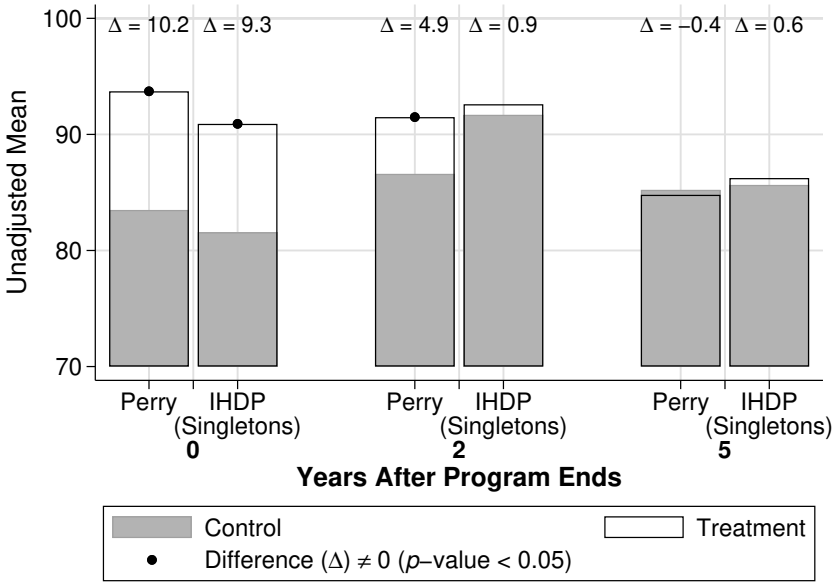
(b) Perry and IHDP Twins



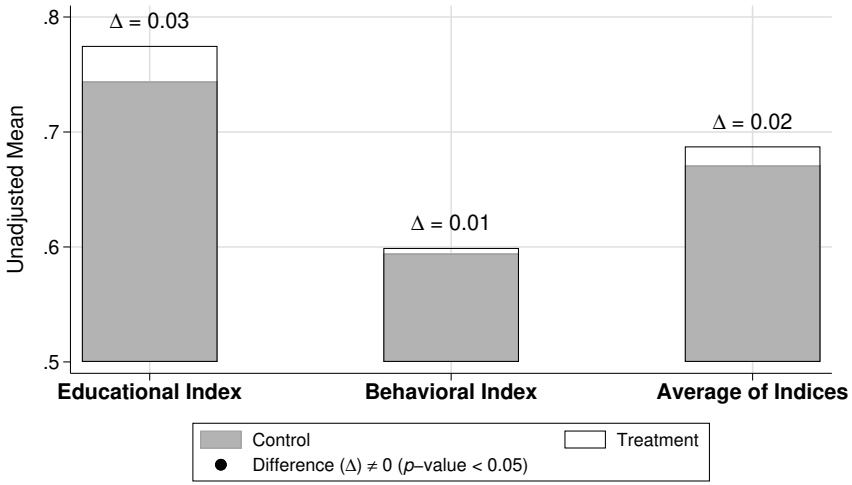
Note: Panel (a) displays the average cognition by treatment status for participants of the Perry Preschool Program and the IHDP singletons. It then displays these same averages for the IHDP singletons by treatment site, labeled with the postal abbreviation of the state in which the site is located. For both programs, cognition is measured using the Stanford-Binet IQ test, anchored to its national mean of 100 and standard deviation of 15. For Perry, end-of-program is age 5. For IHDP, it is age 3. The plot labels the treatment effect and sample size. It also labels the treatment effect when the null hypothesis that the treatment effect is 0 is rejected using the significance levels indicated. The p-values are based on robust standard errors clustered at the child-participant level. Panel (b) is analogous in format to Panel (a). For IHDP, this panel focuses on twins only.

Figure 3. Seemingly "Small" Impact of IHDP in the Longer Term

(a) Post-Program Impact on Cognition for Singletons



(b) Age-18 Non-Cognitive Proxies Pooling Singletons and Twins



Note: Panel (a) displays the average cognition by treatment status for participants of the Perry Preschool Program 0, 3, and 5 years after the end of the program. It also displays these averages for the IHDP singletons. For both programs, cognition is measured using tests anchored to their national mean of 100 and standard deviation of 15. For Perry, the test is the Stanford-Binet IQ test for all years. For IHDP, the tests are the Stanford-Binet IQ test (0 years after), the Wechsler PPSI IQ test (2 years after), and the Wechsler ISC IQ test (5 years after). Panel (b) displays the average indices for the age-18 non-cognitive skill proxies by treatment status for the participants of the IHDP. The indices are described in Section 3. The plots label the treatment effect (average treatment-control difference) and whether the null hypothesis that the treatment effect is 0 is rejected using a significance level of 5%. The p-values are based on robust standard errors clustered at the child-participant level.

discussion of the age-18 outcomes and lack of adulthood follow-up of the IHDP participants.

The impacts for the pooled sample in Figure 3b are indeed lukewarm. However, Figure 4 shows that the impacts by twinning are substantial. We provide the impact on the average of the two indices in Figure 3b. For singletons, the average impact is 0.03. IHDP increases by an average of 0.03 each educational and behavioral item (the control-group mean of the average of the indices is 0.67). The impact of 0.03 differs from 0 when using a significance level of 5%; it amounts to 1/5 of the standard deviation of the average of the indices. This result indicates that IHDP scales the success of similar small-scale programs up to age 18.

When inspecting the age-18 results by sex, we find that IHDP increases the average of the academic and behavioral indices by almost 1/3 of the standard deviation when focusing on female singleton participants. Female singleton participants drive the impact on the pooled sample of male and female singleton participants. This result is sensible given that the overall impact is driven by impacts on educational outcomes (see Appendix Table A.8), which, as mentioned above, early childhood education programs affect the most for female participants. The impact on twins is negative and sizable. It differs statistically from 0 when using a significance level of 5%. Though twins represent a small fraction of the sample, knowing why the impact on them is negative is necessary for a complete understanding of IHPD. Before exploring the sources of the impact differences between singletons and twins, we compare the findings in this section to previous evaluation studies of the IHDP.

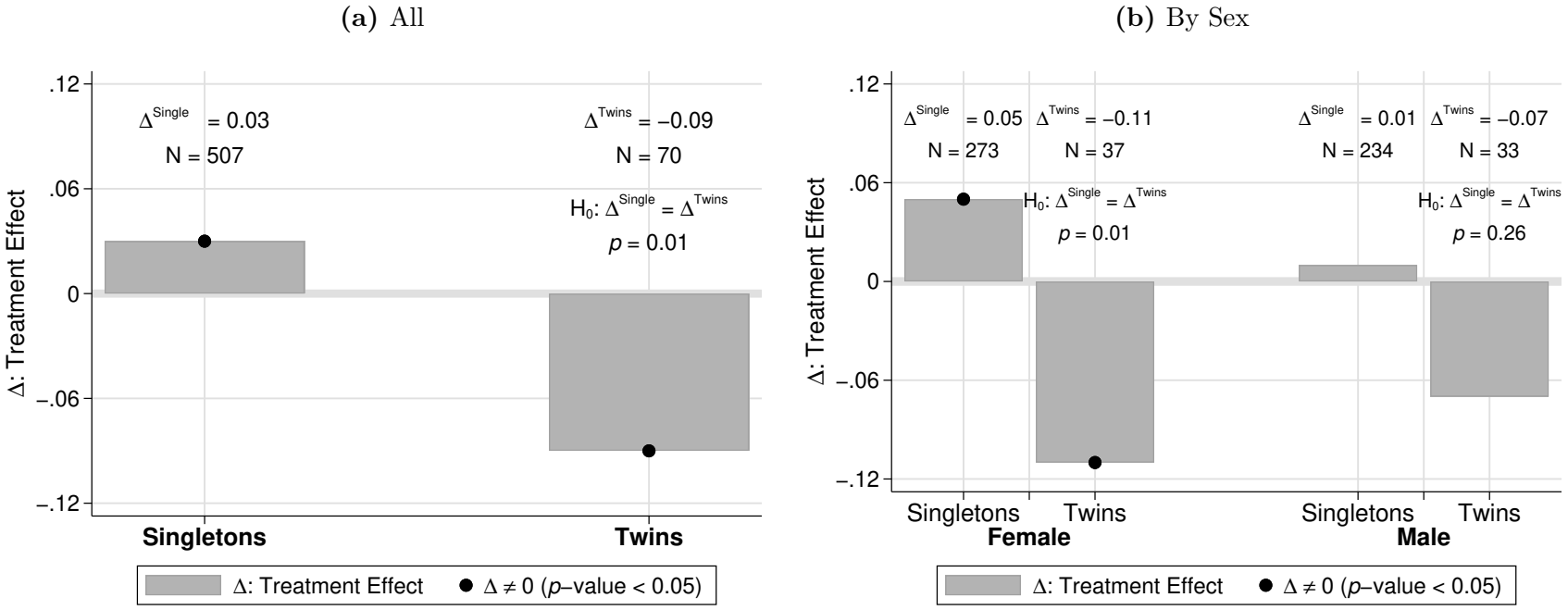
4.3 Related Literature

Our study is not new in evaluating IHDP. There is a literature in the field of Child Development doing so.[19] The evaluations up to age 8 focus on impacts on cognition (e.g., Brooks-Gunn et al., 1992; Liaw et al., 1995; McCormick et al., 2012, 1998). They report an impact on age-3 cognition of at most 3/5 of the national standard deviation, which fades after age 3 and disappears by age 8. These studies pool singletons and twins. The impact they report is relatively small when compared to the impact that we report for singletons.[20]

[19]Our paper also relates to literature documenting a positive impact (e.g., better test scores in elementary and middle school, greater college enrollment) of early life intervention on (very) low-birthweight children (e.g., Bharadwaj et al., 2013; Chyn et al., 2021), who have more health difficulties and worse long-term outcomes relative to peers born with normal birthweight (Almond and Currie, 2011; Currie, 2011).

[20]Other related papers are Duncan and Sojourner (2013) and Chaparro et al. (2020). They both focus on (short-term) cognitive outcomes. Duncan and Sojourner (2013) use the impacts of IHDP on cognition to predict how gaps between different socioeconomic groups would be closed upon a national implementation of the program. Chaparro et al. (2020) combine the experimental data of IHDP with non-experimental data to estimate a household model of decisions regarding child development to evaluate several family policies related to childcare. They focus on outcomes up to age 3. They drop twins from their main sample. Exploration of treatment effects by twinning status or age-18 outcomes is outside of the scope of their study.

Figure 4. Impact on Age-18 Indices of Non-Cognitive Proxies by Twinning Status



Note: Panel (a) displays the treatment effect (average treatment-control difference) on the average of the two indices of age-18 non-cognitive proxies (i.e., the average of the index of educational outcomes and the index of behavioral outcomes), denoted by Δ, for the participants of IHDP. The indices are described in Section 3. The plot labels the treatment effect and sample size. It also labels the treatment effect when the null hypothesis that the treatment effect is 0 is rejected using a significance level of 5%. Additionally, it displays the p-value for the null hypothesis that the treatment effect for singletons equals the treatment effect for twins. The p-values are based on robust standard errors clustered at the child-participant level. Panel (b) is analogous in format to Panel (a) by sex.

Subsequent evaluations study the impact on maternal outcomes (Brooks-Gunn et al., 1994; Martin et al., 2008). They find that, after giving birth to participant children, treatment-group mothers returned to the labor force earlier than control-group mothers and were more likely to be employed. Their likelihood of having children in addition to the participant children did not change relative to the likelihood of control-group mothers, nor did their education. All these findings are based on the follow-ups collected when child participants were three years old or younger. When the child participants were 18 years old, measures of stress and conflict in relationships within the household were collected from the mothers. These surveys had large item non-response rates and no average treatment-control difference in them.

Evaluation studies using the age-18 follow-up with children participants are scarce, with McCormick et al. (2006) being an exception. They find that the IHDP had an impact on performance in some subsections of the Woodcock Johnson Test of Achievement, which measures standardized mathematics and reading knowledge. They also find some modest and imprecisely estimated improvements on a self-reported behavior checklist.[21] They do not study the age-18 outcomes that we analyze. Another exception is Petitclerc and Brooks-Gunn (2022), who evaluate the impact on three outcomes measuring engagement with the criminal justice system at age 18 and find reduced engagement by male participants. We do not have access to the outcomes they analyzed, but we study several other outcomes that these authors do not explore.[22]

Our study is novel in some of its program-evaluation elements. It is the first to provide results by twinning status and to provide an economic justification for doing so.[23] Twinning status turns out be crucial in understanding treatment effects, as parents of singletons and twins differ fundamentally in their response to the IHDP. The difference helps explain why, in the long term, the program is beneficial for singletons but harmful for twins. Our study is also new in studying age-18 outcomes approximating non-cognitive skills, which in turn cause long-term education, earnings, crime, and health outcomes. The evaluation of these outcomes

[21] We do not analyze that behavior checklist. Instead, we focus on concrete educational and behavioral outcomes, which, generally, are more accurate than self reports (Almlund et al., 2011).

[22] Their item non-response for age-18 outcomes is similar to ours. They do not provide analysis by twinning status or link their results to parenting.

[23] Previous evaluation studies of IHDP display program impacts by birthweight, according to the two categories described in Section 2 (low-low birthright and low-high birthright). This exercise is not based on an economic motivation. It is based on the randomization stratification. The difference in the program impacts by twinning status that we document is not confounded by birthweight. First, the OLS and IPW estimators control for birthweight. Second, there is virtually no correlation between twinning and the arbitrary low-low or low-high birthweight category in the sample. The R^2 of a regression of an indicator of being low-high birthweight on a twinning indicator is 0.006.

is thus relevant for a complete assessment of the efficacy of IHDP. The decomposition of the treatment effects into experimentally induced changes in childcare and parenting is another new element. We build towards it next.

5. Childcare and Parenting: Crowd-In or Crowd-Out?

Assignment to treatment provided free childcare of the highest quality for children in the treatment group. By the law of demand, treatment should increase the number of hours children spent in childcare. Figure 5 displays the average number of hours per week spent in childcare by treatment and twinning status. We first average the four periods observed within each child. We then construct the corresponding averages across children of the within-child averages. Recall that, essentially, treatment-group children exclusively attended IHDP childcare centers. We conclude that the take-up difference between the treatment and control groups is entirely driven by random assignment to treatment. We assume henceforth that treatment-control average differences approximate the comparison of IHDP childcare against no childcare hours at all. This assumption allows us to abstract from the (unobserved) quality of childcare attended by the control group.[24] Two facts justify this. First, control-group singletons and twins spent a small average number of hours in childcare. Second, treatment-group singletons spent, on average, 390 percent more hours in childcare than their control-group counterparts. For twins, the corresponding difference is 863 percent.

The impact of treatment assignment on parenting is not straightforward to predict. Figure 5 summarizes our parenting measures, observed at ages 1 and 3. Similarly to the case of the childcare measure, we first average the two observations within all children and then construct the averages across children by treatment and twinning status.[25] For singletons, treatment *increases* average parenting. For twins, it *decreases* average parenting.[26]

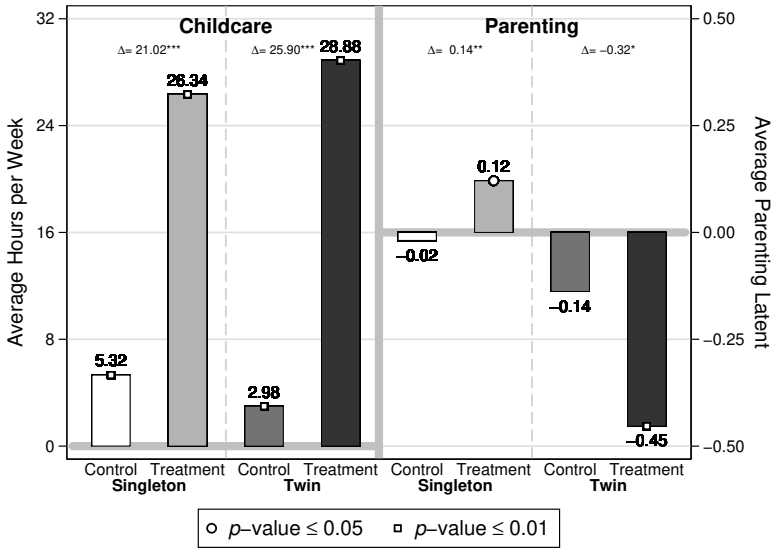
Figure 6 explores raw and adjusted mean differences and provides inference. The specification based on the broadest control set includes child (sex, birthweight, gestational age, race), mother (race, age, education, and employment), and household (poverty status and composition) characteristics, as well as site fixed effects. Further, the treatment effects arise from comparing individuals who are, on average, identical in observed and unobserved

[24] In other contexts, control-group parents enroll their children in alternatives to treatment childcare. In those cases, authors do not focus on the intensive margin (number of hours) but on the discrete choice of enrolling their children in childcare (García et al., 2018; Kline and Walters, 2016).

[25] Appendix Figure A.4 shows that the average across children of the childcare and parenting measures observed at each age follow a very similar pattern as the summary across ages of Figure 5.

[26] Appendix Figure A.2 shows that the treatment-control difference for each of the subscales of the HOME score is qualitatively similar to the treatment-control difference for the aggregate parenting measure, which is based on the subscales. A single scale does not drive the parenting results in this section.

Figure 5. Childcare and Parenting by Treatment and Twinning Status



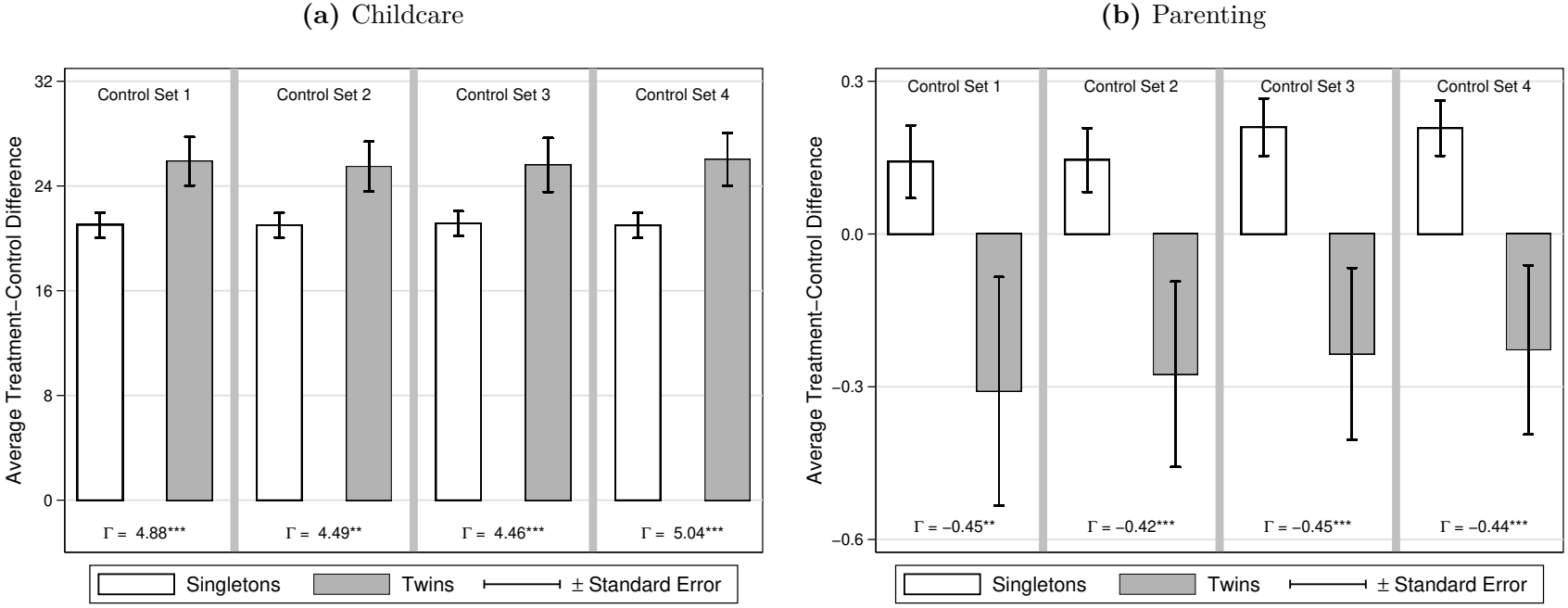
Note: The panel labeled "Childcare" displays the average hours per week in childcare by treatment and twinning status. We first average the observations of average hours per week at ages 18, 24, 30, and 36 months within children. We then average across children to construct the averages for each group in the label. The panel labeled "Parenting" is analogous in format to the panel labeled "Childcare" for the parenting measures observed at ages 1 and 3. The parenting measure displayed is standardized to an in-sample mean of 0 and a standard deviation of 1. $\boldsymbol{\Delta}$ is defined as the treatment-control difference in the corresponding averages. We label averages and average differences according to the p-value associated with their t-statistic. The null hypothesis for either the averages or average differences is that they are 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.

characteristics at baseline. Section 6 thus pursues an explanation for the difference in the treatment effect on parenting that is not based on parental preferences for consumption and child welfare or on household production characteristics other than those related to the possibility that parents of twins supply parenting jointly to their children. Before that explanation, we relate the findings in this section to the current literature.

5.1 Related Literature

This paper is the first to document that high-quality early childhood education can either increase or decrease direct parental involvement depending on factors related to tight birth spacing. Some of the studies discussed in Section 4.3 document sources of treatment-effect heterogeneity, but no study finds a clear switch from parental investment crowd-in to crowd-out when high-quality childcare is provided and birth spacing tightens. In higher levels of education, the crowding-out of parental investment by external provision of educational services has been discussed (e.g., Peltzman, 1973; Pop-Eleches and Urquiola, 2013).

Figure 6. Treatment-Control Difference in Childcare and Parenting by Twinning Status



Note: Panel (a) displays the average treatment-control difference in the mean hours spent in childcare when children were 18, 24, 30, and 36 months old for singletons and twins. To obtain these average differences we regress mean hours spent in childcare on a constant, a twin indicator, a treatment indicator, an interaction of the twin and treatment indicators, and a control set. The average treatment-control difference for the singletons is the estimated coefficient on the treatment indicator. For the twins, it is the estimated coefficient on the treatment indicator plus the estimated coefficient on the interaction. Control Set 1 is empty. Control Set 2 includes the variables in Panel *a* of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels *b* and *c* of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. The standard errors are clustered at the child-participant level. Panel (b) is analogous in format to Panel (a) for the mean of the parenting measures observed at ages 1 and 3. $\mathbf{\Gamma}$ is defined as the twin-singleton difference in the corresponding average treatment effects. We label the estimates of $\mathbf{\Gamma}$ according to the p-value of the t-statistic associated with the difference. The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.

This paper adds to the existing literature demonstrating that parental investments vary as a function of children's prenatal or postnatal characteristics. Almond et al. (2018) and Heckman and Mosso (2014) provide extensive discussions and surveys. Most of these works find that parents remediate unfavorable early life conditions and complement other investments.[27] Salient findings include the following. Parents respond to iodine supplementation for Tanzanian newborns by increasing the vaccination and breastfeeding rates of their children (Adhvaryu and Nyshadham, 2016). Parents remediate early life gender gaps at school entry in Canada, the United States, and the United Kingdom by spending more time with boys than with girls (Baker and Milligan, 2016). Parents in China reallocate their investment in health and education services when one of their children suffers an early life health shock. They focus their investment on health rather than education services for children who suffer the shock. For their other, relatively healthy children, these parents focus on investing in education services (Yi et al., 2015).[28] A related literature indicates that early-life endowments and parental investments are static and dynamic complements in the production of child skills (e.g., Cunha and Heckman, 2008; Cunha et al., 2010; Todd and Wolpin, 2003).[29]

Previous studies exploiting randomized assignment to high-quality early childcare have found that parental investments increase upon this provision. Gelber and Isen (2013) analyze the Head Start Impact Study, a randomized trial of Head Start. They find that Head Start increases parental investment (e.g., time spent reading and doing math with children), and document that these impacts last beyond the intervention period. García et al. (2018) and Appendix J of Heckman and Mosso (2014) document that random assignment to high-quality early childhood education increases parental investment and improves parental practices. The former is based on a program that did not directly treat parents. The latter is based on a program that, like the IHDP, treated parents during home visits, in addition to providing center-based childcare to the children.

6. A Price-Theoretic Model for Interpreting Treatment-Effect Heterogeneity by Twinning Status

The IHDP's 100 percent subsidy to childcare represents a decidedly non-marginal change in the input price ratio for the treated group of parents. As indicated in Figure 6, the response

[27] Almond and Mazumder (2013) present another extensive discussion and survey. They focus on analyzing how parental investments remediate unfavorable prenatal conditions.

[28] Breining et al. (2022) also find that, when one child has a very low birthweight and receives medical treatment, parents invest more in this child's siblings, which indicates complementarity with investments in other children within the household and not only with investments in the "targeted" child.

[29] Related studies are Aizer and Cunha (2012), Attanasio et al. (2020), and Houtenville and Conway (2008).

of those parents is to increase their usage of childcare by an average of more than 20 hours per week. An explicitly discrete analysis of the magnitude of parents' responses (in terms of parenting, labor supply, and consumption) to such a non-marginal change in the cost of childcare would be ideal, but is difficult to perform without assuming specific functional forms for parents' utility and child welfare production functions. A marginal analysis cannot explain the ultimate magnitudes of parents' responses, but it can offer predictions of the direction of those responses with minimal and plausible assumptions about functional forms. In this section we present such analysis, which we use to explore plausible sources of differences in the direction of the effect of a subsidy to childcare on the provision of parenting to singletons and to twins.

We model parents as deriving utility from a numeraire consumption good x, leisure l, and child welfare W. Child welfare, in turn, is determined by parenting within the home (p) and childcare provided externally (c). Parental utility can therefore be represented by the function

$$U\left(x, l, W_r\left(c, p\right)\right), \tag{4}$$

where $W_r\left(c, p\right)$ is the household production function for child welfare and the subscript r takes on either of two values: $r = s$ for a parent of a singleton and $r = t$ for a parent of twins. The utility and production functions are assumed to be strictly concave and homothetic, respectively. At a given wage ω, the parental budget constraint is

$$\omega \cdot \left(H - l_r - p_r\right) = x_r + \tau c_r, \tag{5}$$

where H is the time endowment and τ is the price of a unit of childcare.[30] Parents maximize (4) subject to (5).

In the case of twins, W_r is an aggregation of the welfare levels of her twin children so $W_t := f_t\left(W_1, W_2\right)$, where W_ℓ is the welfare of twin $\ell = 1, 2$. In principle, the levels of childcare and parenting chosen by parents of twins could differ between the two members of the pair. However, it is likely to be quite costly for parents to deviate from equal provision of either input. For childcare, the time and resource cost of transporting each twin to sessions of unequal duration would be greater than the cost of transporting them jointly. In the case of parenting, it would take considerable effort to provide parenting at significantly different

[30] As explained in Section 3, the measure of parenting in our empirical work is expressed in units of effective parenting rather than time. Our analysis in this section expresses parenting in units of time for simplicity. A function linking time to effective parenting would clutter our notation and provide little additional insight.

levels to each twin. For these reasons, we assume that identical levels of the inputs are provided to each member of the twin pair. A household-level joint amount of parenting decided by a parent of twins is enjoyed by her two children. These assumptions allow us to bypass factors such as preferential treatment or utility complementarity within children of the same family (e.g., Behrman et al., 1982).[31]

We let childcare be equally productive and equally provided to each child in a twin pair. Further, we let childcare be equally productive for a singleton child than for each child in a twin pair. That is, let $c_t = c_t^1 + c_t^2$ be the sum of childcare units provided to twins and consider the point $p_s = p_t$ and $c_t^1 = c_t^2 = c_s$. At this point, $\frac{\partial W_t}{\partial c} = 2\frac{\partial W_s}{\partial c}$. To model the joint supply of parenting to twins, we allow the production function of the welfare of twins to differ from the production function of the welfare of singletons by a scalar γ with $1 < \gamma < 2$ so that, at the comparison point, $\frac{\partial W_t}{\partial p} = \gamma\frac{\partial W_s}{\partial p}$ (if the production functions are homethetic, the equality holds at all welfare levels with the same c to p ratio).[32] A unit of childcare provides a marginal welfare gain of $\frac{\partial W_s}{\partial c}$ to a parent of twins, while a unit of parenting provides her with a marginal welfare gain of $\gamma\frac{\partial W_s}{\partial p}$.

The properties of the production functions of welfare indicate that for $p_s = p_t$ and $c_t^1 = c_t^2 = c_s$, the marginal rates of substitution between childcare and parenting for a parent of singletons and a parent of twins are $\frac{\frac{\partial W_s}{\partial c_s}}{\frac{\partial W_s}{\partial p_s}}$ and $\frac{2}{\gamma}\frac{\frac{\partial W_s}{\partial c_s}}{\frac{\partial W_s}{\partial p_s}}$. The price ratio confronting both types of parents is $\frac{\tau}{\omega}$ and their marginal rates of substitution need to equal this ratio at their optima. Therefore, the childcare-parenting ratio provided by a parent of singletons is too high to be optimal for a parent of twins if $1 < \gamma < 2$ (i.e., if parenting is jointly supplied for twins). At the respective optima, $\frac{c_s}{p_s} > \frac{c_t}{p_t}$, and the share of costs devoted to childcare by a parent of singletons relative to the share devoted to parenting is greater than the comparable ratio for a parent of twins.[33] Since the sum of the two cost shares for each parent type is one, a lower relative cost share devoted to formal childcare corresponds to a lower absolute share as well. The evidence on actual cost shares presented in Section 5 indicates that $\gamma < 2$,[34]

[31]Recall that our preferred results in Section 5 are conditional on child-specific endowments that vary within twin pairs such as birthweight. By accounting for these endowments, we abstract from preferential treatment and compensatory investment by parents within their twin children as in Yi et al. (2015).

[32]γ could also help summarize fruitful interactions between a twin pair or human-capital spillovers (e.g., one child teaches another while interacting with their parent), but that element is likely to be minor because the children are between ages 0 and 3 in our empirical analysis of the joint determination of c and p.

[33]The full cost of childcare includes costs that do not vary with the number of hours of care per day, such as transportation to and from the site. We focus on variable costs. As described in Section 2, the subsidy covers the full transportation cost for either the singleton child or the twin children (i.e., the subsidy provided by the IHDP is the same for both singletons and twins). In addition, transportation costs are small relative to the variable cost of each hour of childcare, especially because children who lived far from the treatment sites were disqualified from program participation.

[34]Precisely, for fixed and equal input prices, p and c should be greater for twins than for singletons.

and we assume throughout the rest of this section that this condition is met.

The first-order conditions of the parental optimization problem yield the factor demands for c and p as functions of W, τ, and ω. We use the demand for p to study the impact of IHDP, which we conceptualize as a subsidy on τ. We analyze the parental response to the subsidy focusing on the variable cost shares by decomposing its impact on parenting into local scale and substitution effects, using a hat to denote percentage changes in variables.[35]

At the initial values of c and p (e.g., at the control-group averages of c and p), the optimal marginal percentage change in parenting upon the subsidy is

$$\widehat{p}_r = \varepsilon_{pW_r} \cdot \widehat{W}_r + \theta_{cr} \cdot \sigma_r^{\text{prod}} \cdot \widehat{\tau}, \tag{6}$$

where ε_{pW_r} is the elasticity of the demand of input p with respect to the level of welfare chosen by the parent when relative input prices are unchanged, θ_{cr} is the share of production cost allocated to childcare, σ_r^{prod} is the elasticity of substitution between parenting and childcare in the production of welfare, and $\widehat{\tau}$ is the incremental percentage change in the cost of childcare due to the subsidy (assignment to treatment).

The parental demand for overall child welfare is a function of her income and the cost of such welfare. The effect of a change in that cost is

$$\widehat{W}_r = \varepsilon_{WC_r} \cdot \widehat{C}_r, \tag{7}$$

where ε_{WC_r} is the elasticity of demand for child welfare and C_r is the parental cost of

We examine Figure 5 for control-group participants to verify this implication. An average control twin pair receives 2×3.0 hours of childcare, while an average control singleton receives 5.3 hours. Appendix Figure 5 indicates that this verification also holds for parenting. An average control twin pair receives $2 \times 48.1 = 96.2$ units of parenting, while an average control singleton receives 50.8 units. Additionally, Appendix Figure 5 indicates that the inequality holds for control-group children. For an average control-group twin pair, $\frac{p_t}{c_t} = \frac{2 \times 48.1}{2 \times 3} = \frac{48.1}{3.0} = 16.0 > 9.6 = \frac{50.8}{5.3} = \frac{p_s}{c_s}$. That is, $\frac{p_t}{c_t} \div \frac{p_s}{c_s} \approx 1.6$. Two additional aspects are worth noting. First, comparing control-group twins and singletons indicates that $\frac{p_t}{2} = 48.1 < 50.8 = p_s$, which hints at joint supply of parenting. Second, the large substitution away from parenting by parents of twins upon assignment to treatment is only plausible if p_t is large, relative to p_s, which the data indicate is true. At the control-group average: $p_t = 96.2 > 1.9 \times 50.8 \approx 1.9 \times p_s$. Another empirical verification of the mechanisms proposed in this section is the following. In Section 4, we document that female participants drive the positive (singletons) and negative (twins) longer-term impacts. Female participants should therefore drive the childcare-parenting differential impacts by twinning status. Appendix Figure A.5 provides this verification.

[35] Note that our comparative statics are in elasticity form; thus, the cross effect in production between c and p is contained in the elasticity of substitution.

producing a unit of welfare. By the envelope theorem,

$$\widehat{C}_r = \theta_{pr} \cdot \widehat{\omega}_r + \theta_{cr} \cdot \widehat{\tau} \tag{8}$$

for a given wage rate ω and where θ_{pr} is defined analogously to θ_{cr}. Further, the Slutsky decomposition of ε_{WC_r} is

$$\varepsilon_{WC_r} = \varepsilon^{\text{comp}}_{WC_r} - \theta_{IC_r} \cdot \theta_{cr} \cdot \eta, \tag{9}$$

where $\varepsilon^{\text{comp}}_{WC_r}$ is the compensated version of ε_{WC_r}, θ_{IC_r} is the parental share of full income devoted to child welfare, and η is the income elasticity of demand for child welfare.

Now, note that $\varepsilon^{\text{comp}}_{WC_r} = -(1-\theta_{IC_r}) \cdot \sigma^{\text{cons}}$, where σ^{cons} is the elasticity of substitution in consumption (i.e., the elasticity of substitution between x and W in the parent's consumption). We are interested in demand differences that are not driven by preferences across types r so we do not index either σ^{cons} or η. Substituting Equations (7) to (9) into Equation (6) yields

$$\widehat{p}_r = -\theta_{cr} \left\{ \varepsilon_{pW_r} \left[\underbrace{(1-\theta_{IC_r}) \cdot \sigma^{\text{cons}}}_{\substack{\text{substitution effect} \\ \text{in consumption}}} + \underbrace{\theta_{IC_r} \cdot \theta_{cr} \cdot \eta}_{\substack{\text{income effect} \\ \text{on child welfare level}}} \right] - \underbrace{\sigma^{\text{prod}}_r}_{\substack{\text{substitution effect} \\ \text{in production}}} \right\} \widehat{\tau}. \tag{10}$$

The change in parenting generated by the subsidy has three components: 1) Substitution toward child welfare and away from parent's consumption due to the reduction in the relative price of child welfare; 2) The income effect of a reduced cost per unit of child welfare; and 3) A reduction in the optimal parenting to childcare ratio in the production of child welfare. Because $\widehat{\tau}$ is negative, the first two effects are unambiguously positive, and the third effect is unambiguously negative.

The subsidy increases parenting for singletons ($r = s$) and decreases it for twins ($r = t$) if the overall scale effect of the childcare subsidy on the optimal level of child welfare is sufficiently larger for singletons than for twins. Specifically,

$$\sigma^{\text{prod}}_s < \varepsilon_{pW_s} \cdot [(1-\theta_{IC_s}) \cdot \sigma^{\text{cons}} + \theta_{IC_s} \cdot \theta_{cs} \cdot \eta] \tag{11}$$

and

$$\sigma^{\text{prod}}_t > \varepsilon_{pW_t} \cdot [(1-\theta_{IC_t}) \cdot \sigma^{\text{cons}} + \theta_{IC_t} \cdot \theta_{ct} \cdot \eta]\,. \tag{12}$$

A natural point of comparison is $\sigma_t^{\text{prod}} = \sigma_s^{\text{prod}}$. Sufficient conditions for the inequalities to hold in this scenario are $\varepsilon_{pW_t} \leq \varepsilon_{pW_s}$, $\theta_{IC_t} \geq \theta_{IC_s}$, and $\theta_{IC_t} \cdot \theta_{ct} \leq \theta_{IC_s} \cdot \theta_{cs}$, with at least one of the inequalities being strict, and the magnitudes of the inequalities being large enough to generate the necessary overall inequality with respect to σ_s^{prod} and σ_t^{prod}. If $\sigma_t^{\text{prod}} > (<)\ \sigma_s^{\text{prod}}$, the range of elasticities and cost shares for which the two inequalities hold expands (contracts).

The first of these conditions is likely to hold if parenting is jointly supplied to twins. The second of these conditions requires that a parent of twins does not devote a smaller share of her full income to child welfare than a parent of a singleton, which is likely to hold, since even in the presence of joint supply of p, raising two children simultaneously is likely to require more time and monetary resources than raising one child does. The third of these conditions requires that the share of *full* income devoted to external childcare is not larger for a parent of twins than it is for a parent of a singleton, which is also likely to hold. Relative to the share of full income devoted to parenting two children simultaneously, the share devoted to childcare should be small, especially when compared to the share of full income devoted to parenting by a parent of a singleton. We have thus exemplified a scenario where the treatment effects summarized in Figure 6 could emerge.

Temporal Horizon of Model Predictions. The model explains when the treatment effect on parenting can be positive for singletons and negative for twins. However, the model can only rationalize a positive net impact on child welfare for both subsamples, because random assignment to treatment reduces the cost of attaining any level of child welfare. Figure 3 is consistent with this implication. While the program is in place, its impact on welfare, as measured by cognition, is positive for both singletons and twins. However, the impact is smaller for twins, likely because of the parenting crowd-out.

Our model cannot distinguish between short-term and longer-term outcomes once the program ends and childcare is no longer subsidized.[36] However, we speculate that the longer-term negative impact on twins is a consequence of a sustained reduction in parental investment. That is, we conjecture that a parent of treated twins maintains her reduced level of parenting after the program. She thus does not act as if she is aware that parenting has an important effect on long-run outcomes for her children. Recent literature documenting that parents underestimate the productivity of their investment on their children's skill development supports this explanation (Attanasio et al., 2019; Cunha et al., 2013).

A reduction in parental investment could be naturally accompanied by an increase in

[36]The data available are not well-suited to understand the longer-term dynamics of the relevant choices.

labor force participation. The choice of p embeds a time-allocation decision by a parent that includes time spent in the labor force. Appendix Figure A.6 shows that, while deciding to spend less time and fewer resources on their children relative to their singleton-parent counterparts, parents of treated twins spend more time in the labor force.[37] Our empirical measure of p combines both time and resources (see Section 3.2). If a parent works more, she may be able to spend less time with her children but, at the same time, devote more material resources to them. However, our results indicate that the overall impact on p is negative for a parent of treated twins. Below, we argue that the decrease in p mediates a long-term negative impact on measures of non-cognitive skills, which, in turn, affect outcomes such as labor income, criminal activity, and health. A cost-benefit analysis for twins would likely indicate negative impacts on these non-cognitive outcomes, along with a positive impact on labor incomes of the mothers.

7. Childcare and Parenting as Determinants of Skills: Treatment-Effect Decomposition

To quantitatively link the program impact on the inputs of the production of child welfare, parenting and childcare, to the program impact on outcomes, we use the framework of Heckman et al. (2013). In a Laspeyres or Oaxaca-Blinder exercise, we decompose the treatment effect on outcomes into the experimentally induced changes in childcare and parenting documented in Section 5. We model the outcome of individual i when assigned to treatment status d as follows:

$$Y_i^d = \alpha_0^d + \alpha_{\mathrm{c}}^d \cdot c_i^d + \alpha_{\mathrm{p}}^d \cdot p_i^d + \upsilon_i^d, \tag{13}$$

where α_0^d, α_c^d, and α_p^d are coefficients, c_i^d and p_i^d are the levels of childcare and parenting when child i is assigned to treatment status d,[38] and υ_i^d is an error term. For now, we assume that the error term is mean-independent. We explain alternatives below.

The decomposition is

$$\underbrace{\mathbb{E}\left[Y_i^1 - Y_i^0 | c_i, p_i\right]}_{\text{ATE}} = \underbrace{\left(\alpha_0^1 - \alpha_0^0\right)}_{\substack{\text{unexplained or} \\ \text{residual}}} + \underbrace{\alpha_{\mathrm{c}} \cdot \left(c_i^1 - c_i^0\right)}_{\text{due to childcare}} + \underbrace{\alpha_{\mathrm{p}} \cdot \left(p_i^1 - p_i^0\right)}_{\text{due to parenting}}, \tag{14}$$

[37] Appendix Figure A.6 displays the average treatment effect on the average hours worked a week by mothers when children are 18, 24, 30, and 36 months (we do not have information for fathers). The variable construction is analogous to that of childcare described in Section 3.2. The difference in the average treatment effect between singletons and twins does not differ statistically from 0; its magnitude is large and robust to the specifications considered for childcare and parenting in the main text.

[38] Correspondingly, the observed variables are $c_i = D_i \cdot c_i^1 + (1 - D_i) \cdot c_i^0$ and $p_i = D_i \cdot p_i^1 + (1 - D_i) \cdot p_i^0$.

where we impose the restrictions $\alpha_c^0 = \alpha_c^1 =: \alpha_c$ and $\alpha_p^0 = \alpha_p^1 =: \alpha_p$.[39] We estimate the decomposition at the average by twinning status so the treatment-control differences in the elements due to childcare and parenting are the ATEs discussed in Section 6. We provide decomposition results for age-3 outcomes only. The item non-response due to attrition in the age-18 outcomes generates a sample that is too small to provide precisely estimated results. After discussing the decomposition, we speculate about the extent to which the conclusions extend to age-18 outcomes. We focus on the Stanford-Binet IQ scores here. The decomposition based on PPVT is very similar (see Appendix Figure A.12).

Appendix Table A.10 displays OLS estimates of α_c and α_p obtained in the pooled sample (singletons and twins of the treatment and control groups). To ease interpretation, we standardize both childcare and parenting to an in-sample mean of 0 and standard deviation of 1. Using the most comprehensive specification and under mean independence, we estimate that a one standard deviation increase in parenting increases the IQ test score by 6.5 points. We also estimate that a one standard deviation increase in childcare increases that score by 3.8 points. Inserting these estimates into the decomposition in Equation (14) indicates the following.[40] For singletons, the marginal effect of parenting is larger than the marginal effect of childcare. However, the larger experimentally induced increase in childcare means that this component explains a greater fraction of the ATE. Parenting explains 1.3 points of the ATE, which amounts to 10 points (see Figure 7), while childcare explains 5 points. The unexplained component amounts to 3.5 points. A remarkable aspect of this decomposition is that two variables explain 65 percent of the entire ATE, which suggests that the program's impact is largely due to the two channels analyzed in this paper.

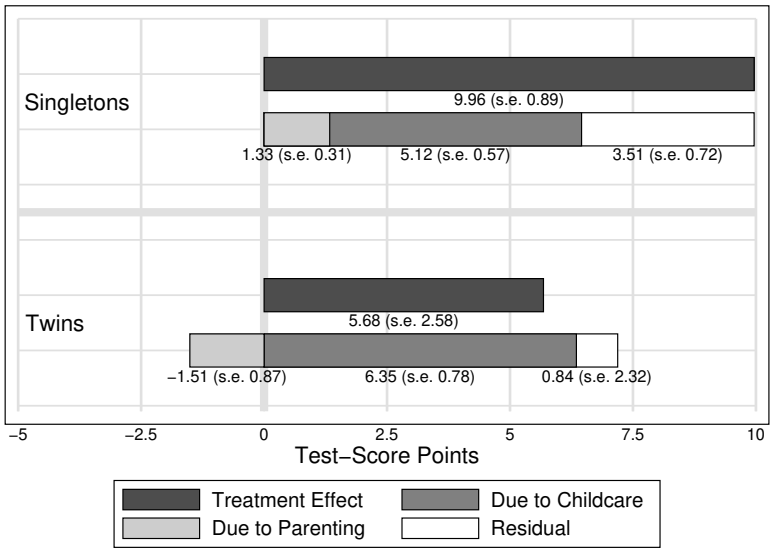
Figure 7 also presents the decomposition of the ATE for twins. This decomposition indicates that, holding parenting constant, the impact of the IHDP amounts to 6.4 points. However, the parenting crowd-out decreases the ATE for twins by 1.5 points, to 5.7 points. The unexplained fraction of the ATE is minor for twins.

Decomposition Based on Instrumental Variables. The mean-independence assumption imposed on the the error term of Equation (13) could be violated if unobserved determinants of the parental decisions underlying childcare and parenting and the realized skills of children are correlated. In this case, the OLS estimates of α_c and α_p would be inconsistent. Appendix 5.2 pursues an alternative strategy based on instrumental variables. In summary,

[39] Appendix Table A.9 indicates alignment of these coefficients across treatment and control groups.

[40] The component $\left(\alpha_0^1 - \alpha_0^0\right)$ in Equation (14) contains the intercept difference in Equation (13). In specifications with controls, it also includes the sum of the marginal effect of each control multiplied by the average difference between the treatment and control groups. The average difference between the treatment and control groups in the control variables is minor by randomization.

Figure 7. Decomposition of Treatment Effect on Age-3 Cognition



Note: This figure displays estimates of the average treatment effect (ATE) decomposition into childcare, parenting, and an unexplained or residual component based on Equation (14). We use the ordinary least squares (OLS) coefficient estimates of α_c (childcare) and α_p (parenting) in Equation (13) reported in Appendix Table A.10. We use the coefficient estimates obtained when using the most comprehensive control set (Control Set 4). The measure of age-3 cognition is the Stanford-Binet IQ. The ATE decomposed is the OLS estimate in Table A.8, also based on the most comprehensive control set.

the IV estimates indicate a larger marginal effect of childcare (α_c) and parenting (α_p) on skills than the OLS estimates do, but the decomposition based on IV provides a qualitatively similar conclusion to the decomposition based on OLS. The decomposition of the program's impact into these two variables is thus robust to the instrumental-variable alternative.

A question that naturally arises is whether the decomposition results apply to the age-18 outcomes. There are three reasons why we speculate that they do. First, there is a qualitative alignment in the program impacts across outcomes regarding their differences between singletons and twins. Singletons benefit from the program, while twins benefit less. In the case of the age-18 outcomes, the program may even harm twins. As explained in Section 6, this fact is consistent with a sustained reduction in parental investment due to misperceptions about the productivity of such investments in the production of skills. Second, the decomposition indicates that parenting and childcare are the major drivers of treatment effects. Thus, if they drive the age-3 outcomes, they plausibly also drive the age-18 outcomes. Third, all successful early childhood education programs impact age-3 cognition in the short term and then long-term non-cognitive outcomes. The mediators of these program impacts do not change over time (Conti et al., 2016; Heckman et al., 2013).

8. Summary and Final Comments

We analyze the Infant Health and Development Program, which aimed to replicate the success of the iconic Perry Preschool and Carolina Abecedarian Projects at scale. Unlike previous analyses of the program, we integrate the non-trivial feature of twin oversampling among program participants into the analysis. We estimate treatment effects by twinning status and study long-term non-cognitive skill proxies not previously analyzed. We find that for singleton participants, who represent the great majority of the sample, the IHDP scales the success of the iconic programs. The IHDP is still a relatively small program if compared, for instance, to the approximately one million children served by Head Start in a given year, but its sample is nine to ten times larger than the sample of Perry or ABC. Our evidence is a first step in documenting that there is no "voltage drop" in the scaling of programs like Perry and ABC.

For singletons, the impact of the program on age-3 cognitive skills is comparable to that of the highest-quality early childhood education programs, but the large initial boost appears to fade out after age 8. However, the impact of the IHDP on their non-cognitive skills is sustained up to age 18. These overall positive impacts are relevant to literature studying the remediation of adverse neonatal conditions (e.g., Almond et al., 2018). For twins, there is a small short-term gain in cognitive skills. More importantly, the program negatively impacts age-18 measures of non-cognitive skills, which is counterintuitive and therefore requires specific analysis.

We document that, upon randomization to treatment, parents of singletons increase both their use of childcare and their own parenting. Parents of twins, in contrast, increase their use of childcare but decrease their parenting (i.e., treatment crowds out parenting for twins). We develop a price-theoretic model to explain how the possibility of jointly supplying parenting to twins and the resulting difference in relative costs of raising tightly spaced children compared to singletons can rationalize this result. This explanation relates our work to classic studies on the consequences of birth spacing on the development of children (e.g., Buckles and Munnich, 2012), birth spacing and parental behavior (e.g., Heckman et al., 1985; Heckman and Walker, 1990), and the potential parental investment crowd-out generated by subsidizing educational services (e.g., Peltzman, 1973). Our findings for the twins suggest that, when birth spacing is tight, subsidizing childcare may be insufficient for generating long-term gains in human-capital. Policies that subsidize multiple inputs of the production function of child skills (e.g., both childcare and parenting) may be necessary under such circumstance.

We decompose the program treatment effects and show that, holding the crowd-out of parenting constant, the childcare component benefits twins to a great extent. The parenting crowd-out does not entirely offset this benefit in the short-term. We speculate that parents sustain the reduction of their own parenting in the long term when the program is no longer in place, generating the negative longer-term impact on twins. That is, the parenting crowd-out generated by the program offsets the positive impact of the program's main component (childcare) for tightly spaced children. This generates negative longer-term consequences, given the apparent sustained crowd-out after the program ends. This is an important policy-design consideration for the US, where the parents of disadvantaged children, who benefit the most from early childhood education, are likely to have tight birth spacing (Gemmill and Lindberg, 2013; Masinter et al., 2017). It is also important for developing countries, where early childhood education programs are growing in prevalence and birth spacing is generally tight (Casterline and Odden, 2016).

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When Should You Adjust Standard Errors for Clustering? *Quarterly Journal of Economics 138*(1), 1–35.

Adhvaryu, A. and A. Nyshadham (2016). Endowments at Birth and Parents' Investments in Children. *Economic Journal 126*(593), 781–820.

Agostinelli, F. and M. Wiswall (2016). Estimating the Technology of Children's Skill Formation. NBER Working Paper w22442, National Bureau of Economic Research.

Aizer, A. and F. Cunha (2012). The Production of Human Capital: Endowments, Investments and Fertility. NBER Working Paper w18429, National Bureau of Economic Research.

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality Psychology and Economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The Costs of Low Birth Weight. *Quarterly Journal of Economics 120*(3), 1031–1083.

Almond, D. and J. Currie (2011). Killing Me Softly: The Fetal Origins Hypothesis. *Journal of Economic Perspectives 25*(3), 153–72.

Almond, D., J. Currie, and V. Duque (2018). Childhood Circumstances and Adult Outcomes: Act II. *Journal of Economic Literature 56*(4), 1360–1446.

Almond, D. and B. Mazumder (2013). Fetal Origins and Parental Responses. *Annual Reviews of Economics 5*(1), 37–56.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2020). Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia. *American Economic Review 110*(1), 48–85.

Attanasio, O., F. Cunha, and P. Jervis (2019). Eliciting Maternal Expectations about the Technology of Cognitive Skill Formation. NBER Working Paper w26516, National Bureau of Economic Research.

Aylward, G. P., S. I. Pfeiffer, A. Wright, and S. J. Verhulst (1989). Outcome Studies of Low Birth Weight Infants Published in the Last Decade: a Metaanalysis. *Journal of Pediatrics 115*(4), 515–520.

Bailey, M. J., S. Sun, and B. Timpe (2021). Prep School for Poor Kids: The Long-run Impacts of Head Start on Human Capital and Economic Self-sufficiency. *American Economic Review 111*(12), 3963–4001.

Baker, M. and K. Milligan (2016). Boy-girl Differences in Parental Time Investments: Evidence from Three Countries. *Journal of Human Capital 10*(4), 399–441.

Behrman, J. R., R. A. Pollak, and P. Taubman (1982). Parental Preferences and Provision for Progeny. *Journal of Political Economy 90*(1), 52–73.

Bharadwaj, P., K. V. Løken, and C. Neilson (2013). Early Life Health Interventions and Academic Achievement. *American Economic Review 103*(5), 1862–91.

Bradley, R. H. and B. M. Caldwell (1984). The HOME Inventory and Family Demographics. *Developmental Psychology 20*(2), 315.

Bradley, R. H., B. M. Caldwell, J. Brisby, M. Magee, L. Whiteside, and S. L. Rock (1992). The HOME Inventory: A New Scale for Families of Pre and Early Adolescent Children with Disabilities. *Research in Developmental Disabilities 13*(4), 313–333.

Breining, S., N. M. Daysal, M. Simonsen, and M. Trandafir (2022). Spillover Effects of Early-Life Medical Interventions. *Review of Economics and Statistics 104*(1), 1–16.

Brooks-Gunn, J., F. R. Liaw, and P. K. Klebanov (1992). Effects of Early Intervention on Cognitive Function of Low Birth Weight Preterm Infants. *Journal of Pediatrics 120*(3), 350–359.

Brooks-Gunn, J., M. C. McCormick, S. Shapiro, A. Benasich, and G. W. Black (1994). The Effects of Early Education Intervention on Maternal Employment, Public Assistance, and Health Insurance: The Infant Health and Development Program. *American Journal of Public Health 84*(6), 924–931.

Buckles, K. S. and E. L. Munnich (2012). Birth Spacing and Sibling Outcomes. *Journal of Human Resources 47*(3), 613–642.

Bureau of the Census (1986). Federal Expenditure by State for Fiscal Years 1985. Website, https://www2.census.gov/library/publications/1986/governments/fes-85.pdf.

Casterline, J. B. and C. Odden (2016). Trends in Inter-Birth Intervals in Developing Countries 1965-2014. *Population and Development Review*, 173–194.

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research 1*(2), 245–276.

Chaparro, J., A. Sojourner, and M. J. Wiswall (2020). Early Childhood Care and Cognitive Development. NBER Working Paper w26813, National Bureau of Economic Research.

Chyn, E., S. Gold, and J. Hastings (2021). The Returns to Early-life Interventions for Very Low Birth Weight Children. *Journal of Health Economics 75*, 102400.

Conti, G., J. J. Heckman, and R. Pinto (2016). The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour. *Economic Journal 126*(596), F28–F65.

Cunha, F., I. Elo, and J. Culhane (2013). Eliciting Maternal Expectations about the Technology of Cognitive Skill Formation. NBER Working Paper w19144, National Bureau of Economic Research.

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

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the Technology of Cognitive and Non-cognitive Skill Formation. *Econometrica 78*(3), 883–931.

Currie, J. (2011). Inequality at Birth: Some Causes and Consequences. *American Economic Review 101*(3), 1–22.

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.

Elango, S., J. L. García, J. J. Heckman, and A. Hojman (2016). Early Childhood Education. In R. A. Moffitt (Ed.), *Economics of Means-Tested Transfer Programs in the United States*, Volume 2, Chapter 4, pp. 235–297. Chicago: University of Chicago Press.

Federal Reserve Bank of St. Louis (2022). Real Median Household Income by State, Annual. Website, https://fred.stlouisfed.org/release/tables?rid=249&eid=259515&od=1985-01-01#.

García, J. L., F. Bennhoff, J. J. Heckman, and D. E. Leaf (2021). The Dynastic Benefits of Early Childhood Education. NBER Working Paper w29004, National Bureau of Economic Research.

García, J. L. and J. J. Heckman (2023). Parenting Promotes Social Mobility Within and Across Generations. *Annual Review of Economics 15*.

García, J. L., J. J. Heckman, D. E. Leaf, and M. J. Prados (2020). Quantifying the Life-cycle Benefits of an Influential Early-childhood Education Program. *Journal of Political Economy 128*(7), 2502–2541.

García, J. L., J. J. Heckman, and A. L. Ziff (2018). Gender Differences in the Benefits of an Influential Early Childhood Program. *European Economic Review 109*, 9–22.

Gelber, A. and A. Isen (2013). Children's Schooling and Parents' Behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics 101*, 25–38.

Gemmill, A. and L. D. Lindberg (2013). Short Interpregnancy Intervals in the United States. *Obstetrics and Gynecology 122*(1), 64–71.

Gray-Lobe, G., P. A. Pathak, and C. R. Walters (2023). The Long-term Effects of Universal Preschool in Boston. *Quarterly Journal of Economics 138*(1), 363–411.

Gross, R. T. et al. (1993). *Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight, Premature Infants in the United States, 1985-1988, Volume 1.* Inter-university Consortium for Political and Social Research [distributor].

Gross, R. T., D. Spiker, and C. W. Haynes (1997). *Helping Low Birthweight, Premature Babies: The Infant Health and Development Program.* Stanford University Press.

Heckman, J., R. Pinto, and P. Savelyev (2013). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review 103*(6), 2052–2086.

Heckman, J. J., V. J. Holtz, and J. R. Walker (1985). New Evidence on the Timing and Spacing of Births. *American Economic Review 75*(2), 179–184.

Heckman, J. J. and G. Karapakula (2021). Using a Satisficing Model of Experimenter Decision-Making to Guide Finite-Sample Inference for Compromised Experiments. *Econometrics Journal 24*(2), C1–C39.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010). The Rate of Return to the HighScope Perry Preschool Program. *Journal of Public Economics 94*(1–2), 114–128.

Heckman, J. J. and S. Mosso (2014). The Economics of Human Development and Social Mobility. *Annual Reviews of Economics 6*(1), 689–733.

Heckman, J. J. and J. R. Walker (1990). The Relationship between Wages and Income and the Timing and Spacing of Births: Evidence from Swedish Longitudinal Data. *Econometrica 56*(8), 1411–1441.

Hojman, A. P. C. (2016). *Three Essays on the Economics of Early Childhood Education Programs.* Ph. D. thesis, The University of Chicago.

Houtenville, A. J. and K. S. Conway (2008). Parental Effort, School Resources, and Student Achievement. *Journal of Human Resources 43*(2), 437–453.

Hoy, E. A., J. M. Bill, and D. H. Sykes (1988). Very Low Birthweight: A Long-term Developmental Impairment? *International Journal of Behavioral Development 11*(1), 37–67.

IPUMS USA (2022). 1980 and 1990 5% Sample. Website, https://usa.ipums.org/usa/sampdesc.shtml#us1990a.

Kline, P. and C. R. Walters (2016). Evaluating Public Programs with Close Substitutes: The Case of HeadStart. *Quarterly Journal of Economics 131*(4), 1795–1848.

Leonard, C. H., R. I. Clyman, R. E. Piecuch, R. P. Juster, R. A. Ballard, and M. B. Behle (1990). Effect of Medical and Social Risk Factors on Outcome of Prematurity and Very Low Birth Weight. *Journal of Pediatrics 116*(4), 620–626.

Leventhal, T., A. Martin, and J. Brooks-Gunn (2004). The EC-HOME across Five National Data Sets in the 3rd to 5th Year of Life. *Parenting 4*(2-3), 161–188.

Liaw, F. R., S. J. Meisels, and J. Brooks-Gunn (1995). The Effects of Experience of Early Intervention on Low Birth Weight, Premature Children: The Infant Health and Development Program. *Early Childhood Research Quarterly 10*(4), 405–431.

Linver, M. R., A. Martin, and J. Brooks-Gunn (2004). Measuring Infants' Home Environment: The IT-HOME for Infants between Birth and 12 Months in Four National Data Sets. *Parenting 4*(2-3), 115–137.

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Currency.

Martin, A., J. Brooks-Gunn, P. Klebanov, S. L. Buka, and M. C. McCormick (2008). Long-term Maternal Effects of Early Childhood Intervention: Findings from the Infant Health and Development Program. *Journal of Applied Developmental Psychology 29*(2), 101–117.

Masinter, L. M., B. Dina, K. Kjerulff, and J. Feinglass (2017). Short Interpregnancy Intervals: Results from the First Baby Study. *Women's Health Issues 27*(4), 426–433.

Matte, T. D., M. Bresnahan, M. D. Begg, and E. Susser (2001). Influence of Variation in Birth Weight within Normal Range and within Sibships on IQ at age 7 Years: Cohort Study. *British Medical Journal 323*(7308), 310–314.

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.

McCormick, M. C., J. Brooks-Gunn, K. Workman-Daniels, J. Turner, and G. J. Peckham (1992). The Health and Developmental Status of Very Low Birth-weight Children at School Age. *Journal of the American Medical Association 267*(16), 2204–2208.

McCormick, M. C., S. Buka, J. Brooks-Gunn, M. Salganik, and W. Mao (2012). Effect of Early Educational Intervention on Younger Siblings: The Infant Health and Development Program. *Archives of Pediatrics and Adolescent Medicine 166*(10), 891–896.

McCormick, M. C., C. McCarton, J. Brooks-Gunn, P. Belt, and R. T. Gross (1998). The Infant Health and Development Program: Interim Summary. *Journal of Developmental and Behavioral Pediatrics*.

Murray, C. (2013). The Shaky Science Behind Obama's Universal Pre-K. American Economic Institute, `https://www.aei.org/articles/the-shaky-science-behind-obamas-universal-pre-k/`. Accessed on June 29, 2023.

Nassar, A. H., I. M. Usta, J. B. Rechdan, T. S. Harb, A. M. Adra, and A. A. Abu-Musa (2003). Pregnancy Outcome in Spontaneous Twins Versus Twins Who Were Conceived through In-vitro Fertilization. *American Journal of Obstetrics and Gynecology 189*(2), 513–518.

National Bureau of Economic Research (2022). Vital Statistics Natality Birth Data. Website, https://www.nber.org/research/data/vital-statistics-natality-birth-data.

National Center for Health Statistics (2022). National Vital Statistics System. Website, https://www.cdc.gov/nchs/nvss/births.htm.

Peltzman, S. (1973). The Effect of Government Subsidies-in-Kind on Private Expenditures: The Case of Higher Education. *Journal of Political Economy 81*(1), 1–27.

Petitclerc, A. and J. Brooks-Gunn (2022). Home Visiting and Early Childhood Education for Reducing Justice System Involvement. *Prevention Science*, 1–14.

Pop-Eleches, C. and M. Urquiola (2013). Going to a Better School: Effects and Behavioral Responses. *American Economic Review 103*(4), 1289–1324.

Quandt, R. E. (1958). The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. *Journal of the American Statistical Association 53*(284), 873–880.

Quandt, R. E. (1972). A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association 67*(338), 306–310.

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.

Ramey, C. T. and F. A. Campbell (1984). Preventive Education for High-risk Children: Cognitive Consequences of the Carolina Abecedarian Project. *American Journal of Mental Deficiency 85*(5).

Richards, M., R. Hardy, D. Kuh, and M. E. Wadsworth (2001). Birth Weight and Cognitive Function in the British 1946 Birth Cohort: Longitudinal Population Based Study. *British Medical Journal 322*(7280), 199–203.

Schweinhart, L. J. et al. (1993). *Significant Benefits: The High/Scope Perry Preschool Study through Age 27. Monographs of the High/Scope Educational Research Foundation, No. 10.* ERIC.

Statista (2022). Average Number of Own Children Under 18 in Families with Children in the United States from 1960 to 2020. Website, https://www.statista.com/statistics/718084/average-number-of-own-children-per-family/.

The Wall Street Journal Editorial Board (2022). The Evidence on 'Free' Pre-K. The Wall Street Journal, https://www.wsj.com/articles/the-evidence-on-free-pre-k-vanderbilt-study-build-back-better-11643656440. Accessed on June 29, 2023.

The White House (2013a). Fact Sheet: President Obama's Plan for Early Education for all Americans. Website, https://obamawhitehouse.archives.gov/the-press-office/2013/02/13/fact-sheet-president-obama-s-plan-early-education-all-americans.

The White House (2013b). Remarks by the President in the State of the Union Address. Website, https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/remarks-president-state-union-address.

The White House (2020). The Build Back Better Framework. Website, https://www.whitehouse.gov/build-back-better/.

The White House (2021). Fact Sheet: The American Families Plan. Website, https://www.whitehouse.gov/briefing-room/statements-releases/2021/04/28/fact-sheet-the-american-families-plan/.

Todd, P. E. and K. I. Wolpin (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal 113*(485), F3–F33.

Todd, P. E. and K. I. Wolpin (2007). The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps. *Journal of Human Capital 1*(1), 91–136.

US Bureau of Labor Statistics (2022). Annual Unemployment Rates by State. Website, https://www.icip.iastate.edu/tables/employment/unemployment-states.

Wang, J. and M. V. Sauer (2006). In-vitro Fertilization (IVF): A Review of Three Decades of Clinical Innovation and Technological Advancement. *Therapeutics and Clinical Risk Management 2*(4), 355.

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., J. T. Bond, and J. T. McNeil (1978). *The Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results Through Fourth Grade*. Ypsilanti, MI: HighScope Press.

Whitehurst, G. J. (2013). Can We Be Hard-Headed About Preschool? A Look at Universal and Targeted Pre-K. Brookings, https://www.brookings.edu/research/can-we-be-hard-headed-about-preschool-a-look-at-universal-and-targeted-pre-k/. Accessed on June 29, 2023.

Yi, J., J. J. Heckman, J. Zhang, and G. Conti (2015). Early Health shocks, Intra-household Resource Allocation and Child Outcomes. *Economic Journal 125*(588), F347–F371.