Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start

Yujian Zhou, Qicheng Zhu, Peidong Zhang

December 19, 2025

1 Introduction

This report is a replication study of Deming (2009), Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. The purpose of this project is not only to reproduce numerical results from the paper, but more importantly to carefully reconstruct the full data and variable construction process that underlies the empirical analysis. A central part of this work is understanding how raw survey data from the CNLSY are transformed into the final samples used in the paper.

In our replication work, we focus on rebuilding the data pipeline that leads to Table 1 and Table 2, and on preparing the data structure required for Table 3. This includes identifying the correct source files, applying the same sample selection rules as in the original study, constructing preschool participation variables, defining pre-treatment covariates, and reshaping the data into both child-level and child-by-age formats. At this stage, the report intentionally avoids discussing regression estimates or treatment effects. Instead, it documents in detail the background of the study and the data work we completed to ensure that later empirical results are based on a correct and transparent foundation.

A key motivation for this detailed documentation is that small differences in sample selection or variable construction can lead to large differences in results, especially in sibling fixed effects designs. By clearly explaining each step of the data construction process, this report makes the replication work reproducible and easier to evaluate.

2 Background and Original Study

2.1 Head Start and Early Childhood Interventions

Head Start is a federally funded early childhood program in the United States that began in the mid-1960s as part of the War on Poverty. The program targets children from low-income families and provides a broad set of services, including preschool education, health screenings, nutritional support, and parental involvement programs. The goal of Head Start is not only to improve early academic skills, but also to support children's overall development and long-term life outcomes.

Compared to smaller and more intensive early childhood programs, such as Perry Preschool or the Abecedarian Project, Head Start operates at a national scale and serves a much more diverse population. This large scale makes it highly relevant for public policy, but it also makes evaluation more difficult. Children who attend Head Start differ systematically from those who do not, particularly in terms of family income, parental education, and early health conditions. As a result, simple comparisons between Head Start participants and non-participants are likely to be biased.

2.2 Deming (2009): Data and Design

Deming (2009) addresses these challenges by using data from the Children of the National Longitudinal Survey of Youth 1979 (CNLSY). The CNLSY links detailed information on children to their mothers in the original NLSY79 survey, making it possible to observe multiple children within the same family. This feature allows the paper to use sibling fixed effects as its main identification strategy.

The sibling fixed effects design compares siblings within the same family who experienced different preschool participation statuses. By doing so, the analysis controls for all family-level characteristics that are constant across siblings, such as permanent income, parental preferences, and many aspects of the home environment. This design greatly reduces selection bias relative to cross-family comparisons and is central to the credibility of the paper's results.

The paper uses two main data structures. For descriptive analysis and balance checks, outcomes and covariates are analyzed at the child level. For the main cognitive outcome analysis, the data are reshaped into a child-by-age panel, where each observation corresponds to a cognitive test taken by a child at a specific age. This panel structure allows the paper to estimate how the effects of Head Start evolve as children grow older.

2.3 Main Findings

Deming (2009) finds that Head Start has positive and statistically significant effects on cognitive test scores at young ages, particularly around ages 5 to 6. However, these effects decline over time and become smaller at older ages, a pattern often described as "fade-out." At the same time, the paper shows that Head Start has substantial positive effects on a wide range of adult outcomes, including educational attainment, earnings, criminal behavior, and health.

These findings suggest that early test score gains may not fully capture the long-term benefits of early childhood interventions. This insight has been influential in later research and motivates careful attention to both short-run and long-run outcomes.

3 Data and Variable Construction

3.1 Data Source and Sample Selection

The data used in this replication come from the CNLSY, which contains longitudinal information on children born to women in the NLSY79 cohort. The raw data include 11,428 children observed over multiple survey waves. My first task was to identify and reproduce the sample selection rules described in Deming (2009).

Following the paper, we restricted the sample to children whose preschool participation status could be observed before school entry. This requires limiting the sample to children who were sufficiently old by 1990. We also excluded observations from the low-income white oversample that was later discontinued in the NLSY. These steps are necessary to match the institutional context of the original study.

For analyses that rely on sibling fixed effects, we further restricted the sample to families with at least two eligible children. This restriction is essential because families with only one child provide no within-family variation for identification. An important part of my replication work was carefully tracking how the sample size changes at each step and ensuring that restrictions were applied at the correct stage of the data pipeline.

We also distinguished clearly between child-level samples and child-by-age samples. Child-level samples are used for descriptive statistics and balance tests, while child-by-age samples are required for cognitive outcome analysis. Applying child-level restrictions too early can severely distort the panel structure, and avoiding this mistake was a key lesson from the replication process.

3.2 Preschool Participation

Preschool participation is categorized into three mutually exclusive groups: Head Start, other preschool, and no preschool. These variables are constructed using maternal reports from the CNLSY. In our replication, we reconstructed these indicators directly from the raw data and verified their consistency with the definitions described in the paper and the original Stata do-files.

In addition to baseline preschool indicators, the paper uses versions of these variables that are suitable for sibling fixed effects analysis. In later regressions, preschool participation is interacted with age group indicators to allow treatment effects to vary across developmental stages. Although these interactions are analyzed later, constructing consistent preschool participation variables at the child level is a necessary prerequisite.

3.3 Covariates

All covariates used in the analysis are pre-treatment variables, meaning they are determined before preschool participation. These include measures of birth outcomes, early health conditions, household composition, parental employment, childcare arrangements, and family income. In my replication, I reconstructed each covariate using the same timing and definitions as in the original study.

A central concern in this process was ensuring that covariates used in sibling fixed effects regressions exhibit within-family variation when appropriate. I verified that variables used in Table 2 align with those later included as controls in Table 3. Missing values were handled following the paper's approach, using sample mean imputation combined with missing-value indicator variables.

3.4 Outcome Variables

The main outcome variables in Deming (2009) are cognitive test scores from the Peabody Picture Vocabulary Test (PPVT) and the PIAT Math and Reading assessments. These tests are observed at multiple ages for each child. In preparation for Table 3, I reshaped the data from a wide child-level format into a long child-by-age format, where each row represents a specific test taken by a child at a given age.

I then constructed a summary index of cognitive skills. This index is created by standardizing each test score and averaging across available tests at each age. Constructing this index required careful attention to the timing of tests, age definitions, and missing data patterns. Although regression results using this index are discussed later, building the outcome variable correctly is a central part of the replication work.

3.5 Descriptive Statistics

Before conducting any regression analysis, I used the reconstructed data to reproduce descriptive patterns similar to those reported in the paper. At the child level, descriptive statistics show that Head Start participants come from more disadvantaged backgrounds compared to non-participants. However, when the sample is restricted to sibling fixed effects families, differences in pre-treatment covariates across preschool categories become much smaller.

This pattern supports the identification strategy used in Deming (2009) and provides evidence that the data construction steps in my replication are consistent with the logic of the original study.

4 Empirical Strategy

4.1 Sibling Fixed Effects Model

We follow Deming (2009) and estimate a sibling fixed effects model of the form

$$Y_{ij} = \alpha + \beta_1 HS_{ij} + \beta_2 PRE_{ij} + X_{ij}'\delta + \gamma_j + \varepsilon_{ij}, \tag{1}$$

where i indexes children and j indexes families. The outcome Y_{ij} is either a standardized test score index or a summary index constructed from multiple outcomes, which are described below. The indicator HS_{ij} equals one if child i attended Head Start, PRE_{ij} equals one if the child attended a non-Head Start preschool, and the omitted category is no preschool. The vector X_{ij} collects pre-treatment covariates, including child gender, age-at-test dummies, test-year dummies, and a rich set of early-life family background characteristics. The term γ_j is a family fixed effect that absorbs all family-level characteristics, such as maternal ability, permanent income, and preferences, and ε_{ij} is an idiosyncratic error term.

In our empirical implementation, we closely replicate the original Stata code. We first reshape the test score data into a panel of children's years, restrict to Deming's estimation sample, and construct age groups for ages 5–6, 7–10, and 11–14. For specifications with fixed effects, we implement the family estimator by demeaning the outcome and then running OLS on the transformed data with standard errors clustered at the family level. All specifications in Tables 2–5 use family-clustered standard errors to account for family correlation in outcomes.

4.2 Outcome Indices

Deming (2009) addresses the problem of multiple hypothesis testing by aggregating conceptually related outcomes into standardized indices. We follow the same approach. For each group of outcomes, we first standardize each component variable to have mean zero and standard deviation one in the estimation sample. When necessary, we multiply variables by -1 so that higher values consistently correspond to more favorable outcomes. We then construct the index as the simple average of these standardized components.

For the cognitive test outcomes that appear in Tables 3 and 4, we work with three component tests: the Peabody Picture Vocabulary Test (PPVT), the PIAT Math test, and the PIAT Reading Recognition test. Within each age group (5–6, 7–10, and 11–14), we standardize the percentile scores of each test and then compute, for each child-year, the mean of the available standardized scores. This average is then standardized again within the age group to produce an age-specific test score index, denoted $Test\_std\_5$–6, $Test\_std\_7$–10, and $Test\_std\_11$–14. We finally stack these into a single outcome variable $Test\_std$ by assigning to each observation the index corresponding to its age group. This constructed index serves as the dependent variable in our Table 3 regressions and in the first four columns of Table 4.

The same logic is applied to non-test outcomes and to long-term young adult outcomes used in Table 4 and Table 5. For example, the long-term index in Table 5 aggregates high school graduation, college attendance, idleness, crime, teen parenthood, and self-reported health into a single standardized measure of adult well-being. In our group replication, some members focus on constructing and estimating these non-test and long-term indices, while my part concentrates on the test score index used in Tables 3 and 4.

4.3 Identification Assumptions

The causal interpretation of β_1 and β_2 relies on a standard "selection-on-observables plus fixed effects" assumption. Formally, we require that

$$\mathbb{E}\big[\varepsilon_{ij} \mid HS_{ij}, PRE_{ij}, X_{ij}, \gamma_j\big] = 0 \tag{2}$$

so that, conditional on family fixed effects and observed pre-treatment covariates, within-family variation in preschool participation is as good as random. Intuitively, among families where siblings differ in their participation in Head Start or other participation in preschool, we assume that the choice of which child attends which program is not systematically related to the unobserved outcome, after controlling for X_{ij}.

The sibling fixed effects γ_j absorb any time-invariant factors shared by siblings, such as maternal ability, background family income, parenting style, or neighborhood characteristics that do not change across children. The rich set of pre-treatment covariates X_{ij} further adjusts for observable, child-specific characteristics measured before preschool enrollment, including birth weight, early health conditions, early income measures, and child-care arrangements from ages 0–3. Under these assumptions, the within-family contrasts between siblings who do and do not attend Head Start identify the average causal effect of Head Start participation relative to no preschool (and likewise for other preschools).

Several threats to identification remain. First, parents can engage in "compensatory" behavior, preferentially sending more disadvantaged siblings to Head Start, or "reinforcing" behavior, concentrating resources on children with higher-ability. Both patterns would violate the conditional independence assumption if they are not fully captured by pre-treatment covariates. Second, differential attrition or missing outcomes according to preschool status could bias estimates if, for example, Head Start participants are more likely to remain on the panel through adolescence and into young adulthood. Finally, measurement error in preschool history or test scores may be correlated with treatment status. Our replication inherits these limitations from the original design. Our contribution is to assess whether the main patterns in Deming's results are robust to independent reconstruction of the data set and estimation in Python.

5 Replication Results

5.1 Replication of Table 1: Family Characteristics

We first replicate Table 1 of Deming (2009), documenting differences in family background characteristics by race and preschool participation.

In the full sample, children who attended Head Start come from substantially more disadvantaged backgrounds than those who attended other preschools or no preschool. For example, Head Start participants have lower permanent income, lower maternal AFQT scores, and lower levels of maternal education on average.

These differences are especially pronounced when comparing Head Start participants to children who attended other preschools. This confirms the presence of strong negative selection into Head Start in cross-family comparisons.

However, once the sample is restricted to the sibling fixed effects subsample, these differences become much smaller. In several cases, differences in means across preschool categories are reduced by more than half, and in some cases they nearly disappear. This pattern holds for both White/Hispanic and Black children.

Table 1: Selected Family and Maternal Characteristics, by Race and Preschool Status

	White / Hispanic			Black			Head Start–None diff. (in SD units)	
	Head Start (1)	Preschool (2)	None (3)	Head Start (4)	Preschool (5)	None (6)	White/Hispanic (7)	Black (8)
Permanent income	26,553 (26,831) [19,555] (19,097)	52,130 (53,483) [34,577] (48,057)	35,592 (35,121) [23,460] (23,552)	24,005 (23,328) [16,103] (15,538)	32,470 (32,934) [21,939] (26,222)	25,980 (25,931) [18,496] (22,440)	-0.39 (-0.36)	-0.11 (-0.12)
Fixed effects subsample	27,560 (27,418) [22,902] (21,820)	41,882 (42,462) [22,403] (23,464)	35,901 (35,216) [23,600] (23,552)	26,010 (25,093) [19,559] (18,845)	28,940 (29,971) [22,853] (29,785)	24,164 (24,030) [16,314] (19,099)	-0.35 (-0.27)	0.11 (0.06)
Mother < high school	0.51 [0.50] (0.50)	0.18 (0.19) [0.38] (0.40)	0.42 (0.45) [0.49] (0.50)	0.33 (0.36) [0.47] (0.48)	0.20 (0.21) [0.40] (0.41)	0.38 (0.40) [0.49] (0.49)	0.18 (0.13)	-0.10 (-0.09)
Fixed effects subsample	0.53 [0.50] (0.50)	0.25 [0.43] (0.44)	0.41 [0.49] (0.49)	0.39 [0.49] (0.49)	0.27 (0.28) [0.45] (0.45)	0.37 (0.36) [0.48] (0.48)	0.24 (0.28)	0.04 (0.07)
Mother some college	0.22 [0.41] (0.42)	0.41 (0.40) [0.49]	0.23 (0.21) [0.42] (0.41)	0.31 (0.30) [0.46]	0.50 (0.49) [0.50] (0.50)	0.28 [0.45]	-0.02 (0.03)	0.07 (0.03)
Fixed effects subsample	0.16 [0.37]	0.31 [0.46]	0.22 [0.41]	0.32 [0.47]	0.42 [0.50] (0.50)	0.30 [0.46]	-0.15 (-0.16)	0.04
Maternal AFQT	-0.44 (-0.42) [0.73] (0.77)	0.23 (0.24) [0.85] (0.84)	-0.21 (-0.22) [0.86] (0.81)	-0.75 (-0.76) [0.49] (0.49)	-0.51 (-0.52) [0.72]	-0.68 (-0.70) [0.60] (0.59)	-0.27 (-0.25)	-0.12 (-0.11)
Fixed effects subsample	-0.48 [0.70] (0.69)	0.02 (0.05) [0.83] (0.84)	-0.20 (-0.21) [0.82] (0.81)	-0.77 (-0.78) [0.48] (0.50)	-0.63 (-0.62) [0.66] (0.68)	-0.76 [0.56]	-0.34 (-0.32)	-0.02 (-0.03)
Grandmother's education	8.53 (8.54) [3.50] (3.41)	10.62 (10.57) [2.92] (3.03)	9.34 (9.42) [3.36] (3.32)	9.71 [2.56] (2.54)	10.88 (10.89) [2.68] (2.67)	9.70 (9.80) [2.87]	-0.24 (-0.23)	0.00 (-0.03)
Fixed effects subsample	8.51 [3.42]	10.09 (10.20) [3.19] (3.16)	9.54 [3.34] (3.35)	9.82 (9.85) [2.59] (2.58)	10.13 (10.17) [2.76] (2.75)	9.98 (10.03) [2.67] (2.66)	-0.31 (-0.34)	-0.06 (-0.07)
Sample size	364 (426)	745 (838)	1,374 (1,648)	415 (461)	249 (276)	551 (637)		
Sample size — FE	229 (265)	315 (359)	510 (591)	206 (229)	144 (163)	259 (292)		

Notes: Cell entries report means, with standard deviations in brackets. The fixed effects subsample consists of families where at least one sibling (but not all) attended Head Start or another preschool. Head Start–None differences are reported in standard deviation units.

The replication of Table 1 provides important validation for the empirical strategy used in Deming (2009). The large differences observed in the full sample highlight why simple comparisons across families are likely to be biased. In contrast, the much smaller differences in the fixed effects subsample suggest that within-family comparisons substantially improve covariate balance.

This finding supports the use of sibling fixed effects as a credible identification strategy. It also confirms that the data construction and sample selection steps in this replication align closely with those in the original study. As a result, the fixed effects subsample used in later analyses is well-suited for estimating the causal effects of Head Start participation.

5.2 Replication of Table 2: Selection into Head Start

We used my cleaned dataset from the Table 1 and Table 2 construction steps and then restricted it to the sibling fixed effects analysis sample. Concretely, this sample keeps families with at least two eligible children and with within-family variation in preschool participation status, so that sibling fixed effects are identified. I also ensured that each covariate is measured before preschool participation (pre-treatment timing), consistent with the paper.

For each covariate listed in the table, I ran a separate regression of the form:

$$X_{if} = \beta_{HS} \nVdash (\text{Head Start}_{if}) + \beta_{Pre} \nVdash (\text{Other Preschool}_{if}) + \alpha_f + \varepsilon_{if}. \tag{3}$$

where i indexes children and f indexes families. The family fixed effect $alpha_f$ absorbs all characteristics shared by siblings (for example, permanent income, parental preferences, and many aspects of the home environment). The omitted group is “no preschool.” Standard errors are clustered at the family level, following the paper.

For each covariate, we report: (1) the estimated coefficient for Head Start and its standard error, (2) the estimated coefficient for other preschool and its standard error, (3) the control mean (the mean of the omitted group) with its uncertainty, (4) the regression sample size for that covariate. We also matched the ordering and row labels to the paper so that my output can be compared line-by-line.

Table 2: Sibling Differences in Pre-Treatment Covariates, by Preschool Status

	Head Start (1)	Other preschool (2)	Control mean (3)	Sample size (4)
Attrited	0.022 (0.013)	-0.008 (0.016)	0.038 [0.192]	1,314
PPVT score, age 3	2.20 (2.24) (4.82)	-7.16* (4.12)	19.90 [11.10]	195
In mother’s HH, 0–3	0.002 (0.006) (0.029)	-0.028 (-0.038) (0.027)	0.899 [0.302]	1,187
Pre-existing health limitation	-0.001 (-0.006) (0.014)	-0.041** (-0.045) (0.018)	0.040 (0.040) [0.197]	1,187
ln(birth weight)	0.048** (0.050) (0.020)	-0.006 (-0.010) (0.017)	4.702 [0.248]	1,226
Very low BW (<3.31 lbs)	-0.022* (0.012)	-0.004 (0.008)	0.021 [0.145]	1,226
ln(income), age 0–3	-0.012 (0.043)	0.043 (0.033)	9.99 [0.72]	1,186
ln(income), age 3	0.011 (0.085)	0.054 (0.064)	9.98 [0.83]	993
Firstborn	0.016 (0.000) (0.055)	-0.124** (0.000) (0.055)	0.419 (0.000) [0.494]	1,251 (0)
Male	0.000 (0.046)	-0.003 (0.046)	0.503 [0.500]	1,251
Age in 2004 (in years)	0.182 (0.298)	-0.433* (0.249)	23.20 [2.88]	1,251
HOME score, age 0–3	1.98 (3.24)	3.07 (4.10)	38.05 [26.25]	427

Table 2 (continued)

	Head Start (1)	Other preschool (2)	Control mean (3)	Sample size (4)
Mom avg hours worked, year before birth	-1.11	2.06	26.03	377
	(3.14)	(1.87)	[12.15]	
Mom avg hours worked, age 0–1	-1.08	1.77	32.52	379
	(3.16)	(1.72)	[11.07]	
Father in HH, 0–3	0.009	-0.003	0.624	739
	(0.034)	(0.023)	[0.450]	
Grandmother in HH, 0–3	-0.003	-0.049***	0.215	1,190
	(0.024)	(0.019)	[0.325]	
Maternal care, age 0–3	0.019	-0.015	0.689	1,244
	(0.019)	(0.022)	[0.405]	
Relative care, age 0–3	-0.007	0.022	0.180	1,244
	(0.019)	(0.019)	[0.335]	
Nonrelative care, age 0–3	-0.012	-0.006	0.131	1,244
	(0.017)	(0.016)	[0.283]	
Breastfed	-0.053** (-0.017)	-0.010	0.333	1,234
	(0.027)	(0.024)	[0.472]	
Mom smoked before birth	-0.012	-0.005	0.392	1,186
	(0.030)	(0.023)	[0.489]	
Mom drank before birth	0.004	0.010	0.080	1,251
	(0.021)	(0.021)	[0.272]	
Regular doctor's visits, age 0–3	0.043	-0.055	0.383	430
	(0.102)	(0.110)	[0.488]	
Ever been to dentist, age 0–3	0.033 (0.016)	0.008 (0.016)	0.303	401
	(0.137)	(0.137)	[0.461]	
Weight change during pregnancy	0.056 (0.105)	-0.168 (-0.178)	29.71	1,146
	(1.181)	(1.139)	[15.34]	
Child illness, age 0–1	0.016 (0.020)	-0.061 (-0.074)	0.320 (0.520)	1,175
	(0.042)	(0.041)	[0.500]	
Premature birth	-0.048	0.007 (0.003)	0.218	1,175
	(0.034)	(0.034)	[0.413]	
Private health insurance, age 0–3	0.093 (0.084)	0.032 (0.033)	0.447	431
	(0.069)	(0.049)	[0.481]	
Medicaid, age 0–3	0.048 (0.045)	-0.006 (-0.008)	0.376	431
	(0.060)	(0.043)	[0.456]	
Pre-treatment index	0.014 (0.032)	0.047 (0.056)	-0.063 (-0.066)	1,251
	(0.061)	(0.055)	[0.987]	

Notes: Main entries report the published coefficients from the paper. Colored parentheses report our replicated estimates (from `table2_results (1).csv`). Red (blue) indicates the replicated value is larger (smaller) than the published value, based on the rounding shown in the table. Standard errors are in parentheses; control means are reported with standard deviations in brackets. Significance levels in the published table: *** 1%, ** 5%, * 10%.

The replicated table shows that most pre-treatment covariates are not strongly related to preschool choice within families once sibling fixed effects are included. In other words, among siblings, Head Start participation is not systematically higher for children with worse pre-treatment health, worse birth outcomes, or different baseline household structure in most rows. This pattern supports the core identification argument in Deming (2009): the sibling fixed effects approach substantially reduces selection bias that would appear in cross-family comparisons.

At the same time, a small number of rows may show non-zero differences that are statistically meaningful (depending on the covariate and sample size). This is expected in practice, because not all within-family selection channels are perfectly eliminated, and measurement noise can also matter. The main conclusion remains that the overall covariate balance within families is much stronger than in the full sample, which increases confidence in using the sibling fixed effects design for estimating program impacts later in the paper.

5.3 Replication of Table 3: Test Score Effects

Table 3 in Deming (2009) estimates the effect of Head Start and other preschools on a cognitive *test score index* constructed from standardized PPVT and PIAT math/reading scores. Treatment effects are allowed to vary by three age groups (5–6, 7–10, and 11–14) by interacting program indicators with age-group dummies. The unit of observation is child-by-age, and standard errors are clustered at the family level; all specifications include controls for gender, first-born status, and age-at-test and survey-year fixed effects.

Table 3: The Effect of Head Start on Cognitive Test Scores

	(1)	(2)	(3)	(4)	(5)
Head Start					
Ages 5–6	-0.025	0.081 (0.091)	0.093 (0.098)	0.131	0.145
	(0.091)	(0.083)	(0.079)	(0.087)	(0.085)
Ages 7–10	-0.116	0.040 (0.047)	0.067 (0.070)	0.116	0.133
	(0.072)	(0.065)	(0.061)	(0.060)	(0.060)
Ages 11–14	-0.201	-0.053 (-0.041)	-0.017 (-0.012)	0.029	0.055
	(0.070)	(0.065)	(0.061)	(0.061)	(0.062)
Other preschools					
Ages 5–6	0.167	0.022 (0.017)	-0.019 (-0.022)	-0.102	-0.079
	(0.083)	(0.082)	(0.078)	(0.084)	(0.085)
Ages 7–10	0.230	0.111 (0.104)	0.087 (0.082)	0.031	0.048
	(0.070)	(0.064)	(0.061)	(0.061)	(0.065)
Ages 11–14	0.182	0.076 (0.069)	0.037 (0.032)	-0.040	-0.022
	(0.072)	(0.068)	(0.065)	(0.066)	(0.069)
p (all age effects equal—Head Start)	0.074	0.111	0.170	0.092	0.151
Pre-treatment covariates	N	Y	Y	N	Y
Sibling fixed effects	N	N	N	Y	Y
Total number of tests	4687	4687	4687	4687	4687
R^2	0.028	0.198	0.269	0.608	0.619
Sample size	1251	1251	1251	1251	1251

Notes: Main estimates report published coefficients. Parentheses show replicated estimates using our Python code. Red (blue) indicates replicated estimates larger (smaller) than the published values; cells without parentheses indicate no deviation from the published estimates.

We reproduce Deming's specification sequence column-by-column: (1) baseline OLS with the common controls above; (2) adding the full set of pre-treatment covariates (with mean imputation and missingness indicators as in the paper's approach); (3) adding key SES controls (permanent income, maternal AFQT, and maternal education); (4) including sibling (family) fixed effects without the full covariate set; and (5) the preferred specification combining sibling fixed effects with all pre-treatment covariates. We follow the paper in clustering by family/MotherID and in estimating the same age-group interactions so that coefficients are directly comparable to the published table.

Our replication matches the published results extremely closely. In the preferred specification (column (5)), we recover the published coefficients and standard errors for both Head Start and other preschools across all three age groups. The few discrepancies in our table are small and appear mainly in intermediate specifications (especially columns (2)–(3)), where some coefficients differ by only a few hundredths (e.g., modest shifts in the Head Start coefficients at ages 5–6 and 7–10 and slight changes in the adolescent coefficient). These deviations are not large enough to change any qualitative conclusion about the evolution of estimates across specifications.

Consistent with the paper, Table 3 shows that the Head Start coefficient becomes more positive as richer controls are introduced and, in contrast, the other-preschool coefficient shrinks toward zero as selection is addressed. In the preferred column (5), Deming reports a test score gain of 0.145 standard deviations at ages 5–6, fading to 0.133 at ages 7–10 and 0.055 at ages 11–14, and he notes that the null of equal Head Start effects across the three age groups cannot be rejected (p = 0.151). Our replication supports the same "fade-out" pattern and the same inference: sizeable early gains, partial attenuation by adolescence, and statistically indistinguishable effects across age groups at conventional levels.

5.4 Replication of Table 4: Head Start Overall and by Subgroup Effect

Table 4 extends the preferred specification from Table 3 (family fixed effects plus the full set of pre-treatment covariates) to multiple indices and subgroups. Panel A reports results for the full sample: the first three columns correspond to the age-specific test score index effects (5–6, 7–10, 11–14), column (4) pools ages 5–14 for the test score index, column (5) reports a *nontest* school-age index (grade retention and learning disability), and column (6) reports a *long-term* young adult index (high school graduation, college attendance, idleness, crime, teen parenthood, and health).:contentReference[oaicite:5]index=5 Subgroup panels (race, gender, and maternal AFQT) are obtained by interacting the Head Start treatment with subgroup indicators; standard errors are clustered at the family level.

We estimate the same equation (1) setup used throughout the paper—including sibling fixed effects and the full covariate set from Table 3 column (5)—and then (i) report age-specific and pooled test-score-index effects, (ii) construct the nontest index and long-term index using the outcome lists described in the paper, and (iii) estimate subgroup effects via interactions with subgroup dummies (black vs. white/Hispanic, male vs. female, low vs. high maternal AFQT).

For the test-score columns (1)–(4) and the nontest index column (5), our replicated coefficients match the published estimates essentially exactly across panels. The most noticeable deviations appear in the long-term index (column (6)), where our replication produces somewhat larger effects for several rows (e.g., in the overall sample and some disadvantaged subgroups), while leaving the broad pattern unchanged. Because column (6) depends on how multiple young-adult outcomes are coded, standardized, and averaged—and on the exact handling of missingness across those items—small implementation differences (variable definitions, index standardization sample, or missing-data rules) can mechanically shift the final index and its standard error even when the test-score side matches closely.

Deming emphasizes two central messages in Table 4. First, the test-score gains show fade-out: positive effects at ages 5–6 and 7–10 with a smaller, less precise effect by ages 11–14. Second, the long-term index impact is large relative to what test scores alone would predict: Deming reports a 0.228 standard deviation improvement in young-adult outcomes, which he interprets as roughly one-third of the bottom-quartile-to-median permanent-income gap and about 75% of

Table 4: The Effect of Head Start Overall and by Subgroup

	Test scores				Nontest score	Long term
	5–6	7–10	11–14	5–14	7–14	19+
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Overall						
Head Start	0.145	0.133	0.055	0.101	0.265	0.228 (0.247)
	(0.085)	(0.060)	(0.062)	(0.057)	(0.082)	(0.072)
Other preschools	-0.079	0.048	-0.022	-0.012 (-0.004)	0.172	0.069 (0.073)
	(0.085)	(0.065)	(0.069)	(0.062)	(0.088)	(0.072)
p (HS = preschool)	0.021	0.254	0.315	0.118	0.372	0.080
Panel B: By race						
Head Start (black)	0.287	0.127	0.031	0.107	0.351	0.237 (0.277)
	(0.095)	(0.075)	(0.076)	(0.072)	(0.120)	(0.103)
Head Start (white/Hispanic)	-0.057	0.111	0.156	0.110	0.177	0.224 (0.215)
	(0.120)	(0.092)	(0.095)	(0.090)	(0.111)	(0.102)
p (black = nonblack)	0.024	0.883	0.317	0.982	0.282	0.924
Panel C: By gender						
Head Start (male)	0.154	0.181	0.141	0.159	0.390	0.182 (0.187)
	(0.107)	(0.079)	(0.081)	(0.076)	(0.123)	(0.103)
Head Start (female)	0.128	0.059	0.033	0.055	0.146	0.272 (0.304)
	(0.106)	(0.083)	(0.085)	(0.081)	(0.108)	(0.106)
p (male = female)	0.860	0.280	0.345	0.346	0.135	0.553
Panel D: By maternal AFQT score						
Head Start (AFQT ≤ -1)	0.171	0.016	-0.023	0.015	0.529	0.279 (0.325)
($n = 361$)	(0.129)	(0.095)	(0.102)	(0.094)	(0.156)	(0.114)
Head Start (AFQT > -1)	0.133	0.172	0.144	0.154	0.124	0.202 (0.205)
($n = 890$)	(0.094)	(0.073)	(0.074)	(0.071)	(0.091)	(0.091)
p (low = high AFQT)	0.809	0.198	0.192	0.245	0.024	0.595

Notes: Standard errors in parentheses.
All results use the specification in column (5) of Table 4, including family fixed effects and pre-treatment covariates.
Main estimates report published coefficients. Parentheses show replicated estimates using our Python code. Red (blue) indicates replicated estimates larger (smaller) than the published values; cells without parentheses indicate no deviation from the published estimates.

the black–white outcome gap in the sample. Subgroup patterns reinforce this point: for African American children, initial test-score gains are large but fade sharply by adolescence, yet long-term gains remain substantial: for boys, test-score improvements are concentrated among males; and for children of low-AFQT mothers, test-score gains fade nearly completely by ages 7–10 while nontest and long-term indices remain large. Our replication mirrors these qualitative conclusions, and (if anything) our slightly larger long-term index estimates in column (6) strengthen Deming's key takeaway that early test-score effects are an incomplete proxy for the program's longer-run benefits.

5.5 Replication of Table 5: Point Estimates for Individual Non-test and Long-term Outcome

In Table 4, columns (5) and (6) report pooled effects of Head Start on two composite measures: a school-age non-test outcome index and a long-term young-adult outcome index. These indices aggregate grade retention, learning disability, high school graduation, college attendance, idleness, crime, teen parenthood, and health into standardized summary measures. While these pooled indices are useful for addressing multiple-testing concerns and for summarizing overall program impacts, they mask potentially important heterogeneity across individual components. Table 5 provides a direct decomposition of the index effects by reporting point estimates for each individual outcome that enters the non-test and long-term indices. The regression specification is identical to the preferred specification used in Table 4, including family fixed effects and the full set of pre-treatment covariates. Subgroup effects are obtained by interacting Head Start participation with indicators for race, gender, and maternal AFQT, allowing the individual components of the pooled effects to be examined across key demographic dimensions. For each outcome and subgroup, the first row reports our replicated coefficient, followed by the corresponding estimate reported in Deming (2009) when it differs from our replication. Standard errors from our replication and from the original paper are reported in parentheses in the third and fourth rows, respectively.

Overall, our replication of Table 5 closely matches the results reported in Deming (2009) across most outcomes and subgroups. For the overall sample, the estimated effects of Head Start on grade repetition, learning disability diagnosis, high school graduation, college attendance, early parenthood, and self-reported health are nearly identical to those in the original paper in both magnitude and statistical significance. Patterns of heterogeneity by race, gender, and maternal AFQT are likewise preserved. As in the original study, non-test educational gains are substantially larger for Black participants and for children of low-AFQT mothers. Gender differences also replicate closely: improvements in college attendance and health are concentrated among females, while reductions in grade retention are driven primarily by males. The large increase in high school graduation among children of low-AFQT mothers—on the order of 17 percentage points—remains one of the most striking findings in our replication.

The primary departures from the paper arise for the idleness and crime outcomes. While the original paper reports a statistically significant reduction in idleness, our replicated estimates are somewhat smaller in magnitude and, in some subgroups, less precisely estimated. Similarly, for crime, our estimates differ modestly from those reported in the paper, though both sets of results point to small and statistically insignificant effects overall. Importantly, these discrepancies do not alter the substantive conclusions: neither the original paper nor our replication finds strong evidence that Head Start has large or robust effects on criminal behavior.

Table 5: Point Estimates for Individual Outcomes

	All	Black	Nonblack	Male	Nonmale	Low AFQT	High AFQT
Grade repetition	−0.072*	−0.110*	−0.030	−0.207***	0.053	−0.144**	−0.033
	(−0.069*)	(−0.107*)	(−0.027)	(−0.204***)	(0.055)	(−0.140**)	(−0.031)
	(0.041)	(0.057)	(0.059)	(0.058)	(0.058)	(0.068)	(0.051)
	(0.040)	(0.056)			(0.057)	(0.069)	(0.050)
Learning disability	−0.058***	−0.070**	−0.046	−0.047	−0.069***	−0.109***	−0.030
	(−0.059***)	(−0.071**)			(−0.070***)		(−0.032)
	(0.021)	(0.028)	(0.030)	(0.031)	(0.026)	(0.043)	(0.021)
				(0.030)		(0.042)	
High school graduation	0.085***	0.109***	0.056	0.114**	0.056	0.167***	0.041
	(0.086***)	(0.111***)	(0.055)		(0.058)		(0.042)
	(0.031)	(0.041)	(0.048)	(0.049)	(0.044)	(0.056)	(0.037)
				(0.048)			(0.036)
not including GED	0.065*	0.067	0.061	0.111**	0.022	0.127**	0.029
	(0.063*)		(0.058)	(0.108**)	(0.021)	(0.126**)	(0.027)
	(0.034)	(0.045)	(0.052)	(0.052)	(0.047)	(0.064)	(0.039)
		(0.044)	(0.051)			(0.063)	(0.038)
At least one year of college attempted	0.056	0.134***	−0.033	0.022	0.090*	0.012	0.080*
	(0.057)	(0.136***)	(−0.034)		(0.091*)		(0.082*)
	(0.036)	(0.049)	(0.051)	(0.046)	(0.055)	(0.051)	(0.047)
			(0.050)	(0.045)	(0.054)		
Idle	−0.096**	−0.068	−0.132**	−0.126**	−0.070	−0.096	−0.083*
	(−0.071*)	(−0.030)	(−0.123**)	(−0.100**)	(−0.043)	(−0.070)	(−0.072)
	(0.039)	(0.057)	(0.057)	(0.053)	(0.053)	(0.075)	(0.047)
	(0.038)	(0.053)	(0.055)	(0.049)	(0.052)	(0.070)	(0.045)
Crime	0.001	0.011	−0.014	0.031	−0.030	−0.004	0.003
	(0.019)	(0.051)	(−0.020)	(0.036)	(0.002)	(0.038)	(0.008)
	(0.039)	(0.050)	(0.058)	(0.059)	(0.053)	(0.069)	(0.046)
	(0.040)		(0.062)	(0.058)	(0.057)	(0.072)	(0.047)
Teen parenthood	−0.020	−0.043	−0.001	0.011	−0.048	−0.038	−0.009
	(−0.019)	(−0.040)			(−0.047)		(−0.008)
	(0.036)	(0.052)	(0.053)	(0.052)	(0.056)	(0.065)	(0.044)
							(0.043)
Poor health	−0.069***	−0.045	−0.096**	−0.035	−0.102**	−0.092*	−0.058*
	(−0.070***)	(−0.047)	(−0.094**)	(−0.036)		(−0.090*)	(−0.060*)
	(0.027)	(0.035)	(0.043)	(0.037)	(0.042)	(0.047)	(0.034)
	(0.026)						(0.033)

Notes: Entries are Head Start coefficients from sibling fixed-effects regressions of individual non-test and long-term effect. Standard errors (in parentheses) are clustered at the family level. Red (blue) indicates replicated estimates larger (smaller) than the published values; cells without parentheses indicate no deviation from the published estimates. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

6 Robustness Checks

6.1 Alternative Sample Selection

We assess the robustness of our results using several alternative sample selection rules. First, the baseline analysis restricts the sample to individuals who were at least 19 years old by 2004, ensuring that respondents were no longer eligible for Head Start participation at later survey dates. Because some outcomes may be sensitive to age, we tighten this restriction by requiring individuals to be at least 20 years old instead. Second, to address concerns that comparisons between siblings who are very far apart in age may reflect differences in family circumstances or macroeconomic conditions, we restrict the analysis to sibling pairs who are no more than five years apart in age. Third, we cap the maximum age of the sample at 25 years, excluding older respondents for whom pre-treatment information may be less complete. Finally, we adopt a stricter definition of preschool participation by eliminating children who report inconsistent Head Start participation histories and by excluding the small number of children who report less than three months of program participation, replacing the baseline HS2/Pre2 indicators with the more restrictive HS3/Pre3 classification. In Table 6, we present the main results(corresponding to Table 4 Panel A).

The robustness checks based on alternative age restrictions and treatment definitions largely confirm the baseline results. Requiring individuals to be at least 20 years old, capping the sample age at 25, and using a stricter definition of program participation all yield estimates that are very close to the baseline specification, with both the non-test score and long-term outcome indices remaining positive and statistically significant. Although individual point estimates vary somewhat across these specifications, none of these changes alters the qualitative nature of the findings or the statistical significance of the main results. Greater discrepancies emerge when restricting the analysis to siblings no more than five years apart in age; however, this instability is concentrated primarily in the test-score outcomes, while the estimated effect on the long-term outcome index remains statistically significant. A likely explanation is the substantial reduction in sample size under this restriction (436 observations, compared with 1,251 in the baseline sample), which reduces statistical power and increases sampling variability. Overall, this pattern suggests that the long-run impacts of Head Start are robust, whereas short-term test-score estimates are more sensitive to sample size and power considerations.

6.2 Spillover Analysis

A potential concern with the sibling fixed-effects design is within-family spillovers: if Head Start changes parenting practices or the home environment, untreated siblings may benefit indirectly from a treated sibling. Such spillovers would attenuate within-family contrasts and could bias treatment effect estimates toward zero. Following Deming (2009) and the approach used in GTC (2002), we test for spillovers by allowing the treatment effect to vary with birth order.

Specifically, we construct interaction terms between preschool participation and an indicator for first-born status. Let HS_{ij} and Pre_{ij} denote indicators for Head Start and other preschool participation, and let $FirstBorn_{ij}$ equal 1 for first-born children. We then estimate the preferred sibling fixed-effects specification augmented with interactions:

$$Y_{ij} = \alpha + \beta_{HS} HS_{ij} + \beta_{Pre} Pre_{ij} + \gamma_{HS}(HS_{ij} \times FirstBorn_{ij}) + \gamma_{Pre}(Pre_{ij} \times FirstBorn_{ij}) + \delta X_{ij} + \varepsilon_i. \tag{4}$$

Under this parameterization, β_{HS} is the Head Start effect for non–first-born children ($FirstBorn = 0$), while $\beta_{HS} + \gamma_{HS}$ is the effect for first-born children ($FirstBorn = 1$). Evidence of spillovers

Table 6: Robustness Checks: Alternative Sample Selection

	Test scores				Nontest score	Long term
	5–6	7–10	11–14	5–14	7–14	19 +
	(1)	(2)	(3)	(4)	(5)	(6)
Age $\geq$ 20(N=1162)						
Head Start	0.144*	0.127**	0.055	0.096	0.290***	0.266***
	(0.092)	(0.063)	(0.064)	(0.059)	(0.088)	(0.075)
Other preschools	−0.032	0.115*	0.052	0.064	0.159*	0.101
	(0.088)	(0.068)	(0.071)	(0.063)	(0.094)	(0.076)
Sibling gap $\leq$ 5(N=436)						
Head Start	0.041	0.136	0.030	0.076	0.030	0.334**
	(0.141)	(0.122)	(0.122)	(0.113)	(0.205)	(0.165)
Other preschools	−0.058	0.098	0.032	0.044	0.199	0.155
	(0.129)	(0.105)	(0.104)	(0.095)	(0.149)	(0.114)
Age capped at 25(N=994)						
Head Start	0.186**	0.177**	0.089	0.145**	0.172*	0.246**
	(0.090)	(0.072)	(0.073)	(0.068)	(0.105)	(0.099)
Other preschools	−0.088	0.016	−0.074	−0.038	0.148*	0.047
	(0.089)	(0.071)	(0.074)	(0.067)	(0.101)	(0.094)
Strict definition (HS3/Pre3)(N=1196)						
Head Start	0.179**	0.140**	0.076	0.117	0.253***	0.234***
	(0.087)	(0.061)	(0.062)	(0.057)	(0.084)	(0.074)
Other preschools	0.008	0.083	0.019	0.043	0.163*	0.058
	(0.081)	(0.063)	(0.068)	(0.060)	(0.087)	(0.076)

Notes: Each panel re-estimates Table 4, Panel A (Overall) under the indicated alternative sample selection rule. Entries are coefficients with clustered standard errors in parentheses. Significance stars: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

from older to younger siblings would imply systematically larger impacts for later-born children, corresponding to a negative and statistically significant γ_{HS}.

Table 7 reports results for the two summary indices. For both the non-test index and the long-term index, the interaction term $HS \times FirstBorn$ is statistically insignificant, indicating that the effect of Head Start does not differ systematically by birth order. In particular, we do not find evidence that Head Start has larger impacts for later-born children. The estimated Head Start coefficients remain stable after allowing for birth-order interactions. For other preschool, the interaction terms with first-born status are also generally statistically insignificant although the interaction for the long-term index is marginally positive. Overall, our interaction-based test finds limited and inconsistent evidence of spillovers, which aligns with the paper.

Table 7: Spillover Test via Birth-Order Interactions

	Head Start	HS×FirstBorn	Other pre	Pre×FirstBorn
Non-test index				
	0.319	−0.125	0.145	0.081
	(0.111)	(0.200)	(0.128)	(0.203)
		$p = 0.532$		$p = 0.690$
Long-term index				
	0.196	0.147	−0.056	0.332
	(0.104)	(0.173)	(0.108)	(0.191)
		$p = 0.396$		$p = 0.082$

Notes: Coefficients come from the preferred sibling fixed-effects specification augmented with birth-order interactions: $HS_i \times FirstBorn_i$ and $Pre_i \times FirstBorn_i$. Standard errors (in parentheses) are clustered at the family level. Reported p-values test the null that the interaction coefficient equals zero.

7 Re-analysis

7.1 Motivation and Method

The main results in Deming (2009) are identified using a sibling fixed-effects design. While this approach addresses many confounding concerns, treatment assignment to Head Start remains non-random at the child level. To further address selection on observables using methods covered in this course, we re-analyze the main adult outcome using an *inverse probability weighting* (IPW) estimator combined with sibling fixed effects. The estimand of interest is the average effect of Head Start participation (relative to no preschool) on the standardized long-term outcome index.

We first estimate the propensity score

$$e(X_{ij}) = \Pr(T_{ij} = 1 \mid X_{ij}),$$

where T_{ij} indicates Head Start participation for child i in family j, and X_{ij} includes observed pre-treatment covariates such as race, gender, birth order and permanent income indicators. The propensity score is estimated via a logistic regression.

Using the estimated propensity scores, we construct stabilized inverse probability weights,

$$w_{ij} = \begin{cases} \dfrac{\Pr(T=1)}{e(X_{ij})}, & T_{ij} = 1, \\ \dfrac{\Pr(T=0)}{1-e(X_{ij})}, & T_{ij} = 0, \end{cases}$$

and trim extreme weights at the tails of the distribution to mitigate sensitivity to limited overlap. The trimmed weights are well behaved, with a maximum of approximately 2.09.

We then estimate a weighted sibling fixed-effects regression of the form

$$Y_{ij} = \alpha + \beta T_{ij} + \gamma^{\top} Z_{ij} + \mu_j + \varepsilon_{ij},$$

where Y_{ij} is the adult outcome index, Z_{ij} includes within-family covariates (gender and first-born status), μ_j is a mother fixed effect, and standard errors are clustered at the family level. This specification targets within-family contrasts while reweighting observations to improve balance on observed covariates.

7.2 Results and Comparison

Table 8 reports the IPW sibling fixed-effects estimates for the effect of Head Start participation relative to no preschool on the standardized long-term outcome index has point estimate of 0.234 with a standard error 0.076.

Table 8: IPW Sibling Estimates of Head Start on Long Term Outcome

	$\hat{\beta}$	Standard Error	p-value	N
Head Start (vs. none)	0.234	0.076	0.002	887

Table 9 presents corresponding subgroup estimates. The estimated effects are positive across all groups considered, with larger and more precisely estimated effects for Black children ($\hat{\beta} = 0.264$, $p = 0.013$) than for non-Black children ($\hat{\beta} = 0.199$, $p = 0.063$). Gender-specific estimates are also positive, with somewhat larger point estimates for females than for males.

Table 9: IPW Sibling Fixed-Effects Estimates by Subgroup

Subgroup	$\hat{\beta}$	Standard Error	p-value	N	Families
Black	0.264	0.107	0.013	229	85
Non-Black	0.199	0.107	0.063	195	72
Female	0.288	0.156	0.065	130	53
Male	0.188	0.140	0.181	110	50

These patterns closely mirror the long-term results reported in Column (6) of Table 4. In that column, we get a long-term Head Start effect of 0.228 (SE = 0.072) for the full sample, 0.237 (SE = 0.103) for Black children, and 0.182 (SE = 0.103) for males. The magnitudes of our IPW sibling fixed-effects estimates are strikingly similar: the full-sample estimate of 0.234 is nearly identical to Deming's 0.228, and the subgroup estimates for Black and male children align closely with those reported in the original analysis. The close correspondence between the IPW results and original sibling fixed-effects estimates suggests that selection on observed child-level covariates plays a limited role in driving the main long-term findings.

Overall, the IPW re-analysis still provides an important robustness check using methods covered in this course and the re-analysis strengthens the conclusion that Head Start participation generates meaningful improvements in adult outcomes, particularly for historically disadvantaged groups.

8 Discussion and Limitations

Our replication largely reproduces Deming (2009) and confirms the main qualitative patterns: test-score gains are strongest at younger ages and partially fade out by adolescence, while impacts on non-test and long-term indices remain positive and economically meaningful. Our robustness exercises generally support the stability of the long-term results. In addition, the IPW sibling fixed-effects re-analysis yields estimates very close to the baseline sibling fixed-effects specification, suggesting that selection on observed child-level covariates is unlikely to be the primary driver of the long-term findings.

However, there are several limitations remain. First, the sibling fixed-effects design addresses family confounding but does not guarantee identification if parents differentially allocate Head Start to siblings based on unobserved, specific traits. Second, family spillovers could attenuate family contrasts. Although our birth order interaction test finds limited evidence of spillovers, this approach may not capture all channels of indirect effects. Third, parts of the analysis rely on self-reported outcomes and composite indices, which are sensitive to coding choices, standardization samples, and the quality of data. Finally, several robustness restrictions substantially reduce the sample size, weakening test-score outcomes precision and limiting the ability to detect heterogeneous effects.

9 Conclusion

This report independently reconstructs the CNLSY-based data pipeline and replicates the core results of Deming (2009) on the effects of Head Start. Consistent with the original study, we find evidence of positive early test-score impacts that fade over time, alongside robust improvements in non-test and long-term outcomes. Extensions using alternative sample definitions and an IPW sibling fixed-effects approach yield similar estimate results, reinforcing the conclusion that Head Start generates meaningful benefits beyond short-term cognitive test scores. Overall, our replication supports the policy relevance of large scale early childhood interventions and highlights the importance of measuring long-term outcomes when evaluating such programs.

References

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