REPLICATION AND REANALYSIS OF EARLY CHILDHOOD INTERVENTIONS AND LIFE-CYCLE SKILL DEVELOPMENT: EVIDENCE FROM HEAD START [DEMING, 2009]

FINAL PROJECT FOR STAT 256

Zekai Wang
zekai.wang@berkeley.edu

Boyu Fan
david_fan1@berkeley.edu

ABSTRACT

This paper replicates and extends the analysis of [Deming, 2009], which evaluates the long-term impact of the Head Start program on life-cycle skill development. Using the National Longitudinal Mother-Child Supplement dataset, we validate the causal framework, replicate the primary findings, and assess the robustness of results through alternative estimation strategies. Additionally, we test unconfoundedness using Kolmogorov-Smirnov (KS) tests, analyze covariate distribution similarity using kernel density estimates (KDEs), and extend the model with interaction terms to explore heterogeneous effects.

1 Introduction

The Head Start program, launched in 1965 as part of the War on Poverty, is a federally funded pre-school initiative aimed at promoting school readiness among children from low-income families. By providing comprehensive educational, health, and social services, Head Start seeks to bridge early developmental gaps and improve long-term outcomes. Despite serving over 900,000 children annually with a budget of $6.8 billion, skepticism persists about the program's effectiveness, with critics questioning its long-term benefits.

Deming (2009) addresses these concerns by using the National Longitudinal Mother-Child Supplement (CNLSY) dataset to estimate the program's impact on life-cycle skill development. Focusing on a cohort enrolled between 1984 and 1990, Deming employs a within-family fixed-effects framework to control for unobserved heterogeneity. His findings indicate that Head Start participation improves a summary index of young adult outcomes by 0.23 standard deviations, closing one-third of the gap between children in the bottom and median income quartiles. While test score gains fade out by adolescence, the study highlights significant long-term benefits, particularly for disadvantaged subgroups.

The primary research question in the original paper is whether participation in Head Start leads to measurable long-term improvements in life outcomes, and if so, to what extent these benefits persist across different socioeconomic groups. The study also explores the mechanisms driving these outcomes, questioning the reliability of short-term test scores as predictors of broader life-cycle benefits.

Our replication and extension of Deming's analysis aim to assess the robustness of these findings and evaluate the sensitivity of the results to methodological changes. We replicated Table 1 - 5 of [Deming, 2009]. In addition, we conduct robustness checks to test the unconfoundedness assumption by using Kolmogorov-Smirnov (KS) tests on stratified propensity scores and analyze covariate distribution similarity through kernel density estimates. In addition, our reanalysis incorporates interaction terms into the Ordinary Least Squares (OLS) model to capture heterogeneity in treatment effects and employs the Hajek estimator as an alternative method for causal effect estimation. We also use the Doubly Robust estimator that incorporates both inverse propensity weighting and the OLS model. By revisiting the original data cleaning and analytical procedures, we aim to validate the causal framework, replicate key findings, and contribute to the ongoing discourse on early childhood interventions and their long-term impact.

Related Work The effect of Head Start on various outcomes has been widely studied in the literature. For example, [Currie and Thomas, 1995] similarly observes that, although Head Start has a positive effect on test scores, the effect fades out as age grows, especially for African Americans compared to Whites. Head Start reduces the probability of K12 grade repetition for Whites but does not have a significant effect on African Americans. Head Start participants from both races enjoy better long-term health. The results are consistent with Deming [2009] in that we also observe stronger fade out among African Americans and observe that both races enjoy significant long-term benefits.

The Head Start Impact Study conducted by the United States Department of Health and Human Services [**?**] further corroborates these findings. This landmark study employed a randomized controlled trial (RCT) design with a nationally representative sample of over 4,600 children across 23 states. The study demonstrates that Head Start participants experience substantial short-term cognitive gains, including improved test scores, as well as non-cognitive benefits such as reductions in grade repetition and learning disability diagnoses. However, the study also confirms the fade-out effect of test score gains by adolescence. Despite this, the long-term impacts on participants remain substantial, encompassing higher educational attainment, better health outcomes, and reduced involvement in crime and teen parenthood.

2 Methodology of Deming [2009]

2.1 Multiple Inference

The study examines the effects of Head Start on several outcomes: the test score outcome aggregates over the PIATMT, PIATRR, and PPVT test; the short-term non-test score outcome aggregates over K12 grade repetition and diagnosis of learning disabilities; and the long-term outcome aggregates over high school graduation, college attendance, idleness, crime, teen parenthood, and health. Given the relatively small sample size (1455) and the large number of outcomes, the issue of multiple inference becomes a significant concern. To address this, the study constructs summary indices, which aggregate multiple outcomes into single measures. This approach is robust to multiple inference because it reduces the likelihood of Type I errors as the number of outcomes increases.

The summary index is constructed by normalizing each outcome to have a mean of zero and a standard deviation of one. This normalization ensures that outcomes are on a comparable scale. The formula for normalization is:

$$\tilde{Y}_{ij} = \frac{Y_{ij} - \bar{Y}_j}{\sigma(Y_j)},$$

where Y_{ij} represents the outcome of individual i for outcome j, $\bar{Y}_j$ is the mean, and $\sigma(Y_j)$ is the standard deviation. The signs of the normalized outcomes are then adjusted so that positive values consistently indicate favorable outcomes. The summary index is calculated as the simple average of the normalized outcomes:

$$I_j = \frac{1}{N} \sum_{i=1}^{N} \tilde{Y}_{ij},$$

where N is the number of outcomes included in the index. Afterward, the index is standardized once more to zero mean and unit variance. These indices are constructed for the test scores, non-test school-age outcomes (e.g., grade retention, learning disability diagnosis), and long-term outcomes. Although information is lost about each individual outcome, by aggregating outcomes into summary indices, the study provides a measure of program impact that is robust to the multiple inference problem.

2.2 OLS with Fixed Effects

Given that Head Start participants come from more disadvantaged backgrounds, simple comparisons of outcomes between Head Start participants and non-participants are likely to be biased. To mitigate this bias, the study employs an OLS model with family fixed-effects. This approach leverages sibling comparisons within families to control for unobserved, time-invariant factors that may influence outcomes, such as permanent income and maternal cognitive ability.

The estimating equation for the family fixed-effects model is:

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

where Y_{ij} represents the outcome of individual i in family j, HS_{ij} is an indicator for Head Start participation, PRE_{ij} is an indicator for other preschool participation, X_{ij} is a vector of family-varying covariates, γ_j is the family fixed effect, and ϵ_i is the individual-specific error term. The fixed effect γ_j accounts for unobservable family-level factors,

ensuring that comparisons are made within families. The key identifying assumption is that within-family variation in pre-school participation is uncorrelated with unobserved determinants of the outcomes:

$$E(\epsilon_i|X_{ij}, HS_{ij}, PRE_{ij}, \gamma_j) = 0.$$

This assumption implies that, conditional on the included covariates and family fixed effects, the assignment of siblings to Head Start or other pre-schools is effectively random. Another key assumption is ignorability, which we defer to Section 4 for detailed discussion.

The family fixed effects provide the model with great flexibility. However, it still assumes a parametric model between the treatment, the covariates, the family fixed effects, and the outcomes. We discuss the use of alternative causal effect estimators in Section 5.

3 Replication

3.1 Dataset

The dataset for this replication study is derived from the Child and Young Adult cohort of the National Longitudinal Survey of Youth (CNLSY). The CNLSY began in 1986 and tracks the children of the original NLSY women respondents. These children provide a rich longitudinal dataset with information on cognitive, educational, and socioeconomic outcomes. The CNLSY offers unique advantages for evaluating Head Start due to its detailed sibling information and comprehensive maternal data collected before the birth of the children.

For this replication, we restrict the sample to children who were age-eligible for Head Start by 1990 and age 19 or older by 2004. This ensures that all participants completed the Head Start eligibility period and are old enough for long-term impact evaluation. After applying the age restriction and excluding oversampled low-income white respondents (removed from the NLSY after 1990), the dataset includes 3,698 children from families with at least two age-eligible siblings.

The CNLSY tracks Head Start participation via maternal responses collected biannually from 1988 onward. Mothers reported whether their children had ever participated in Head Start or other pre-school programs. These responses enable sibling comparisons within the same family to identify the causal effects of Head Start. Additionally, the dataset includes extensive maternal background data, such as Armed Forces Qualification Test (AFQT) scores, educational attainment, work history, and income, which serve as crucial pre-treatment covariates.

3.2 Data Cleaning

Eligibility Criteria. Several restrictions were applied to identify children eligible for Head Start participation. First, children included in the analysis must have been age-eligible for Head Start by 1990, ensuring they had completed the eligibility period. Additionally, families were required to have at least two age-eligible children, allowing for within-family comparisons of Head Start participation and facilitating the fixed-effects analytical framework. Deceased children and families lacking sibling variation in program participation were excluded to maintain the validity of the sibling-based comparisons and prevent bias in the estimates.

Handling Missing Data. Missing data was addressed through mean imputation (conditioned on the same race and gender of the child) and indicator variables for imputed responses. For example, missing age values were filled using information from adjacent survey years, minimizing the loss of observations due to incomplete records. Program participation indicators for Head Start, other pre-school programs, and no pre-school were constructed based on maternal survey responses collected biannually from 1988 onward. Within-family variation in treatment was captured by constructing indicators for sibling differences in program participation, ensuring the analysis appropriately addressed the study's causal framework.

Program Participation. Program participation indicators were constructed to identify children who attended Head Start, other pre-schools, or no pre-school. These variables were based on maternal responses collected in biannual surveys from 1988 onward. To facilitate fixed-effects analysis, additional indicators were created to capture within-family variation in program participation. This included flags for sibling differences in treatment, ensuring the analysis was robust to nonrandom assignment at the family level.

Covariate Construction. We constructed and standardized covariates to represent the summary index mentioned in the 2: permanent family income, maternal education, and maternal AFQT scores were constructed and standardized to ensure consistency across observations. Family income data was adjusted to 2004 dollars using the Consumer

Price Index, and the average income across all survey years was used to compute a measure of permanent income. Maternal education was categorized into three levels—high school dropout, high school graduate, and some college education—offering a detailed view of the family's educational background. Maternal AFQT scores were adjusted for age at the time of testing and standardized to facilitate cross-family comparisons. For observations with missing AFQT scores, conditional means were imputed based on maternal age at childbirth and racial background to preserve data completeness while ensuring robust covariate construction.

Table 1: SELECTED FAMILY AND MATERNAL CHARACTERISTICS, BY RACE AND PRE-SCHOOL 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	26,388	50,042	35,154	22,789	32,405	25,211	-0.29	-0.11
(Std Dev)	[19,459]	[45,940]	[23,424]	[14,835]	[26,157]	[21,756]		
Fixed Effects Subsample	26,575	45,533	36,482	23,876	30,637	23,698	-0.40	0.01
(Std Dev)	[21,132]	[25,011]	[24,515]	[16,325]	[26,976]	[18,692]		
Mother < High School	0.52	0.20	0.44	0.36	0.23	0.40	0.17	-0.08
(Std Dev)	[0.50]	[0.40]	[0.50]	[0.48]	[0.42]	[0.49]		
Fixed Effects Subsample	0.50	0.21	0.38	0.40	0.26	0.38	0.25	0.04
(Std Dev)	[0.50]	[0.41]	[0.49]	[0.49]	[0.44]	[0.49]		
Mother Some College	0.22	0.39	0.23	0.29	0.48	0.29	-0.02	0.01
(Std Dev)	[0.41]	[0.49]	[0.42]	[0.45]	[0.50]	[0.45]		
Fixed Effects Subsample	0.19	0.37	0.26	0.29	0.42	0.31	-0.17	-0.04
(Std Dev)	[0.39]	[0.48]	[0.44]	[0.45]	[0.50]	[0.46]		
Maternal AFQT (Std)	-0.46	0.16	-0.25	-0.76	-0.51	-0.68	-0.25	-0.13
(Std Dev)	[0.72]	[0.85]	[0.85]	[0.50]	[0.71]	[0.61]		
Fixed Effects Subsample	-0.47	0.08	-0.22	-0.78	-0.58	-0.71	-0.30	-0.11
(Std Dev)	[0.68]	[0.86]	[0.82]	[0.50]	[0.69]	[0.59]		
Grandmother's Education	8.48	10.57	9.23	9.68	10.80	9.99	-0.22	-0.11
(Std Dev)	[3.49]	[2.94]	[3.54]	[2.56]	[2.66]	[2.76]		
Fixed Effects Subsample	8.54	10.43	9.49	9.79	10.41	10.17	-0.28	-0.15
(Std Dev)	[3.45]	[3.10]	[3.46]	[2.57]	[2.72]	[2.53]		
Sample Size	421	769	2167	450	261	929		
Fixed Effects Subsample Size	313	490	906	307	199	515		

Summary Statistics. We calculated mean and standard deviation values for important variables within each pre-school status group. Table 1 presents a descriptive overview of family and maternal characteristics stratified by race and pre-school participation status. The table reflects substantial differences in socioeconomic background among White/Hispanic and Black families across the Head Start, other pre-school, and no pre-school groups. For instance, Black Head Start participants show lower maternal AFQT scores and permanent income compared to their White/Hispanic counterparts, indicating the program's targeted approach toward economically disadvantaged families. These differences underscore the importance of accounting for heterogeneity in treatment effects, as outcomes are likely influenced by baseline disparities in socioeconomic status. We also provide summary statistics for the outcomes and covariates, as well as their interpretations, in Table 12 in the Appendix.

3.3 Replication of Main Results

The replication of the main results focused on assessing the impact of Head Start participation on cognitive test scores and noncognitive outcomes using a sibling fixed-effects framework. By comparing within-family variations, the analysis aimed to eliminate biases due to unobserved family-level characteristics, ensuring a robust estimation of the effects of the pre-school programs.

3.3.1 Examination of Pre-treatment Covariates.

The examination of pre-treatment covariates, as summarized in Table 2, revealed no significant differences between children enrolled in Head Start, other pre-school programs, and those with no pre-school participation. These findings are consistent with random variations rather than systematic selection into different pre-school programs, supporting the validity of the fixed-effects framework. Notably, differences in covariates such as birth weight, maternal work history, and family structure were minimal and were unlikely to introduce significant biases into the analysis. The inclusion of imputed mean values for missing data ensured that the sample size was maintained and attrition rates were low.

Table 3 presents the estimated effects of Head Start on cognitive test scores across three age groups: early childhood (ages 5–6), primary school (ages 7–10), and adolescence (ages 11–14). In the initial model without pre-treatment covariates (Column 1), the estimated effect of Head Start on test scores for ages 5–6 was negative, at -0.063, and further declined to -0.230 for ages 11–14. However, as covariates and fixed effects were progressively included, the estimated effects improved substantially.

Table 2: SIBLING DIFFERENCES IN PRE-TREATMENT COVARIATES, BY PRE-SCHOOL STATUS

	Head Start (1)	**Other Preschool (2)**	**Control Mean (3)**	**Sample Size (4)**
Attrited	0.022 (0.018)	-0.006 (0.022)	0.041 [0.198]	1517
PPVT Score, Age 3	-1.252 (21.10)	-6.891 (18.13)	20.72 [12.94]	261
ln (Birth Weight)	0.048 (0.027)	-0.006 (0.023)	4.71 [0.253]	1424
Very Low BW (<3.31 lbs)	-0.022 (0.017)	-0.004 (0.010)	0.017 [0.129]	1424
In Mother's HH, 0–3	0.008 (0.054)	0.008 (0.051)	0.689 [0.463]	1384
Pre-Existing Health Limitation	0.000 (0.020)	-0.037 (0.026)	0.042 [0.200]	1384
Firstborn	0.016 (0.075)	-0.124 (0.074)	0.423 [0.494]	1455
Male	0.000 (0.062)	-0.003 (0.063)	0.491 [0.500]	1455
Age in 2004 (Years)	0.182 (0.404)	-0.433 (0.338)	22.82 [2.99]	1455
HOME Score, Age 3	1.983 (7.26)	3.073 (9.18)	40.47 [27.37]	530
Father in HH, 0–3	0.009 (0.058)	-0.003 (0.038)	0.635 [0.447]	887
Grandmother in HH, 0–3	-0.003 (0.033)	-0.049 (0.026)	0.210 [0.330]	1387
Maternal Care, Age 0–3	0.019 (0.026)	-0.015 (0.030)	0.643 [0.413]	1448
Relative Care, Age 0–3	-0.007 (0.026)	0.022 (0.026)	0.197 [0.337]	1448
Nonrelative Care, Age 0–3	-0.012 (0.023)	-0.006 (0.022)	0.161 [0.300]	1448
Breastfed	-0.053 (0.037)	-0.010 (0.032)	0.344 [0.475]	1436
Regular Doctor's Visits, Age 0–3	0.043 (0.227)	-0.055 (0.245)	0.382 [0.486]	537
Ever Been to Dentist, Age 0–3	0.033 (0.315)	0.008 (0.316)	0.292 [0.455]	504
Weight Change During Pregnancy	0.056 (1.65)	-0.168 (1.59)	30.02 [15.31]	1338
Child Illness, Age 0–1	0.016 (0.058)	-0.061 (0.057)	0.528 [0.499]	1370
Premature Birth	-0.048 (0.046)	0.007 (0.046)	0.210 [0.407]	1372
Private Health Insurance, Age 0–3	0.093 (0.152)	0.032 (0.109)	0.525 [0.482]	538
Medicaid, Age 0–3	0.048 (0.134)	-0.006 (0.096)	0.312 [0.448]	538
ln (Income), Age 0–3	-0.012 (0.058)	0.043 (0.045)	10.05 [0.728]	1386
ln (Income), Age 3	0.011 (0.125)	0.054 (0.093)	9.997 [0.902]	1160
Mom Avg. Hours Worked, Year Before Birth	-1.111 (6.38)	2.063 (3.80)	26.16 [11.97]	475
Mom Avg. Hours Worked, Age 0–1	-1.079 (5.76)	1.769 (3.13)	32.08 [11.97]	478
Mom Smoked Before Birth	-0.012 (0.042)	-0.005 (0.032)	0.359 [0.480]	1384
Mom Drank Before Birth	0.055 (0.121)	0.061 (0.090)	0.235 [0.424]	502
Pre-Treatment Index	0.021 (0.080)	0.064 (0.073)	0.000 [1.000]	1455

The R^2 values across columns indicate the significant explanatory power of sibling fixed effects. Without sibling fixed effects (Columns 1–3), R^2 values range from 0.041 to 0.272, while the inclusion of fixed effects (Columns 4 and 5) increased R^2 to 0.613 and 0.623, respectively. These results amplify the importance of accounting for unobserved family-specific factors in evaluating the program's impact. Notably, the inclusion of SES indicators, such as maternal AFQT scores, maternal education, and standardized family income, improved the model's explanatory power but did not diminish the estimated effects of Head Start, suggesting that selection on unobservables aligns with selection on observables.

3.3.2 Subgroup Analysis

The subgroup analysis provides additional insights into the heterogeneity of Head Start's impacts, considering variations across race, gender, and maternal cognitive ability, as summarized in Table 4.

Panel A. Each coefficient in Panel A is derived from an estimation of equation1 using all pre-treatment covariates, as shown in column 5 of Table 3. Test score effects are strongest at ages 5–6 but diminish by adolescence, and this is an indication of the Fade-Out Effects. However, non-test and long-term outcome indices suggest substantial benefits, with Head Start participation leading to a significant improvement of 0.237 standard deviations in long-term outcomes.

Panel B: Race. African American children show large initial test score gains of 0.251 standard deviations, fading by adolescence. Long-term effects remain significant, with gains of 0.273 standard deviations. For Non-Black children, gains are initially small but persist more effectively over time.

Table 3: THE EFFECT OF HEAD START ON COGNITIVE TEST SCORES

	(1)	(2)	(3)	(4)	(5)
Head Start					
Ages 5–6	-0.063 (0.085)	0.060 (0.079)	0.067 (0.075)	0.130 (0.090)	0.146 (0.089)
	[0.463]	[0.442]	[0.373]	[0.150]	[0.101]
Ages 7–10	-0.149 (0.068)	0.004 (0.062)	0.034 (0.059)	0.113 (0.063)	0.128 (0.064)
	[0.029]	[0.945]	[0.563]	[0.074]	[0.045]
Ages 11–14	-0.230 (0.066)	-0.081 (0.061)	-0.048 (0.058)	0.032 (0.064)	0.050 (0.065)
	[0.0005]	[0.184]	[0.408]	[0.618]	[0.440]
Other Preschools					
Ages 5–6	0.296 (0.083)	0.085 (0.077)	0.040 (0.075)	-0.063 (0.086)	-0.035 (0.087)
	[0.0004]	[0.269]	[0.589]	[0.466]	[0.688]
Ages 7–10	0.287 (0.068)	0.103 (0.060)	0.076 (0.057)	0.014 (0.064)	0.034 (0.066)
	[0.00003]	[0.084]	[0.183]	[0.824]	[0.610]
Ages 11–14	0.220 (0.070)	0.055 (0.064)	0.022 (0.062)	-0.042 (0.070)	-0.025 (0.072)
	[0.002]	[0.390]	[0.716]	[0.546]	[0.725]
Permanent income (standardized)			0.128 (0.057)		
			[0.024]		
Maternal AFQT (standardized)			0.296 (0.053)		
			[2.4e-08]		
Mom high school			0.124 (0.065)		
			[0.058]		
Mom some college			0.276 (0.074)		
			[0.0002]		
p (all age effects equal—Head Start)	0.066	0.083	0.128	0.137	0.163
Pre-treatment covariates	N	Y	Y	N	Y
Sibling fixed effects	N	N	N	Y	Y
R^2	0.041	0.214	0.272	0.613	0.623
Sample size	1455	1455	1455	1455	1455

Note: Each cell contains the estimated coefficient with its standard error in parentheses. Below the coefficient and standard error, the p-value for the test of significance is shown in brackets. A smaller p-value indicates greater statistical significance of the result. For example, $p < 0.05$ suggests the effect is statistically significant at the 5% level. Standard errors are adjusted for family-level clustering. Pre-treatment covariates and sibling fixed effects are included as noted in the table.

Panel C: Gender. Males experience less fade-out in test scores than females, with a significant non-test outcome effect of 0.150 standard deviations. Females, on the other hand, show stronger long-term improvements.

Panel D: Maternal AFQT. Low-AFQT mothers' children exhibit significant non-test and long-term gains despite no test score improvement by adolescence. In contrast, children of high-AFQT mothers maintain test score gains and exhibit modest long-term impacts.

3.3.3 Individual Outcomes

Table 5 examines the effects of Head Start participation on a range of individual outcomes. Head Start participants are approximately 6.6 percent more likely to graduate from high school, with significant gains for African American children (9.9 percent) and children of low AFQT mothers (16.5 percent). Excluding GED credentials slightly reduces the effect size.

Participation also reduces the likelihood of being idle by 8.7 percentage points overall, with stronger reductions observed for Non-Black participants (11.8 percentage points) and males (11.0 percentage points). Improvements in college attendance are primarily driven by female participants (14.8 percentage points), with smaller but significant gains observed for the overall sample.

Health outcomes also improve, with Head Start participants 6.8 percentage points less likely to report poor health. This effect is more pronounced for Non-Black participants and children of low AFQT mothers, who experience reductions of 10.0 and 8.7 percentage points, respectively. These results underscore Head Start's effectiveness in promoting educational attainment, reducing disengagement, and enhancing health outcomes, particularly for disadvantaged subgroups.

Table 4: THE EFFECT OF HEAD START OVERALL AND BY SUBGROUP

	Test scores 5–6	Test scores 7–10	Test scores 11–14	Test scores 5–14	Non-test score 5–14	Long term 19+
Panel A: Overall						
Head Start	0.146	0.128	0.050	0.096	0.208	0.237
(SE)	(0.089)	(0.064)	(0.065)	(0.060)	(0.090)	(0.081)
p-value	0.101	0.045	0.440	0.111	0.020	0.003
Other preschools	-0.035	0.034	-0.025	-0.003	0.095	0.072
(SE)	(0.087)	(0.066)	(0.072)	(0.064)	(0.098)	(0.077)
p-value	0.688	0.610	0.725	0.968	0.330	0.348
p (HS = preschool)	0.061	0.228	0.342	0.189	0.321	0.102
Panel B: By race						
Head Start (Black)	0.251	0.143	-0.001	0.095	0.239	0.273
(SE)	(0.109)	(0.079)	(0.078)	(0.075)	(0.122)	(0.114)
p-value	0.021	0.070	0.985	0.205	0.049	0.017
Head Start (NonBlack)	-0.001	0.108	0.127	0.103	0.182	0.204
(SE)	(0.127)	(0.100)	(0.104)	(0.097)	(0.135)	(0.115)
p-value	0.993	0.280	0.226	0.287	0.177	0.076
p (Black = NonBlack)	0.097	0.779	0.318	0.947	0.751	0.670
Panel C: By gender						
Head Start (Male)	0.072	0.116	0.052	0.079	0.150	0.157
(SE)	(0.103)	(0.075)	(0.075)	(0.069)	(0.121)	(0.098)
p-value	0.484	0.120	0.492	0.251	0.216	0.110
Head Start (NonMale)	0.222	0.146	0.053	0.115	0.269	0.309
(SE)	(0.110)	(0.077)	(0.079)	(0.073)	(0.104)	(0.095)
p-value	0.044	0.059	0.500	0.116	0.010	0.001
p (Male = NonMale)	0.201	0.717	0.984	0.638	0.379	0.147
Panel D: By maternal AFQT score						
Head Start (Low AFQT)	0.182	0.040	-0.053	0.013	0.441	0.308
(SE)	(0.137)	(0.100)	(0.103)	(0.097)	(0.173)	(0.132)
p-value	0.183	0.690	0.607	0.892	0.011	0.019
Head Start (NonLow AFQT)	0.139	0.175	0.108	0.141	0.081	0.200
(SE)	(0.105)	(0.079)	(0.082)	(0.076)	(0.104)	(0.103)
p-value	0.183	0.027	0.187	0.064	0.437	0.052
p (Low = NonLow AFQT)	0.787	0.282	0.217	0.301	0.077	0.516
Panel E: p-values for equality of test scores by age group						
	Black	**Nonblack**	**Male**	**Female**	**Low AFQT**	**High AFQT**
p (all effects equal)	0.003	0.240	0.262	0.254	0.198	0.205

4 Robustness Check

[Deming, 2009] conducts various robustness checks. Among other checks, the author varies sample selection criteria by eliminating children with inconsistent responses or altering age restrictions, investigates spill over effects of the treatment, and reestimates the main results with a sampling weight released by the NLSY dataset to make the samples representative of the US population. In this replication, we check by altering the eligibility criteria.

Moreover, in Section **??**, we discussed that [Deming, 2009] adopts the OLS with fixed effects model 1. The model requires two assumptions: ignorability and correct modeling. Ignorability means that the treatment assignment is independent of the potential outcomes given the covariates, i.e., $Y(1), Y(0) \perp T|X$. Correct modeling means that the OLS with fixed effects model is correctly specified. In this section, we check for the ignorability assumption by fitting a propensity score and conducting a covariate balance check. We differ the discussion of the correct modeling assumption to the next section 5.

In the interest of conciseness, we only replicate Panel A of Table 4, which contains the main results of [Deming, 2009]. In the altering eligibility criteria check, we find that although individual point and standard error estimates are sometimes changed, the overall trend is still consistent with the original results. The covariate balance check also shows no evidence of systematic differences in covariate distribution between the treatment and control groups at different strata. These results are in line with [Deming, 2009] and suggest that the results are robust.

For reference, our replicated Panel A of Table 4 is shown below.

4.1 Different Eligibility Criteria

When using age to determine sample eligibility, [Deming, 2009] fills in missing value with ages recorded when administering the Peabody Picture Vocabulary Test (PPVT) test to the child. Instead, we check the robustness of the results by discarding children with missing age values (this is the exact setup of one robustness check in [Deming,

Table 5: POINT ESTIMATES FOR INDIVIDUAL OUTCOMES

	All	Black	NonBlack	Male	Female	Low AFQT	High AFQT
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Grade repetition	-0.042	-0.065	-0.015	-0.084	-0.008	-0.129	0.005
	(0.051)	(0.068)	(0.077)	(0.059)	(0.063)	(0.094)	(0.061)
	[0.415]	[0.338]	[0.841]	[0.155]	[0.893]	[0.169]	[0.929]
Learning disability	-0.061***	-0.066**	-0.058	-0.030	-0.090***	-0.108**	-0.035
	(0.023)	(0.030)	(0.035)	(0.030)	(0.025)	(0.045)	(0.025)
	[0.007]	[0.027]	[0.096]	[0.317]	[0.0002]	[0.016]	[0.157]
High school graduation	0.066**	0.099**	0.031	0.046	0.087**	0.165***	0.013
	(0.034)	(0.044)	(0.054)	(0.044)	(0.040)	(0.060)	(0.041)
	[0.052]	[0.025]	[0.570]	[0.299]	[0.032]	[0.006]	[0.743]
not including GED	0.053	0.056	0.051	0.027	0.077*	0.127*	0.010
	(0.036)	(0.046)	(0.058)	(0.048)	(0.042)	(0.066)	(0.042)
	[0.142]	[0.227]	[0.382]	[0.567]	[0.068]	[0.055]	[0.812]
At least one year of college attempted	0.058	0.117**	-0.006	-0.035	0.148***	-0.005	0.092*
	(0.039)	(0.053)	(0.055)	(0.044)	(0.048)	(0.056)	(0.050)
	[0.133]	[0.029]	[0.917]	[0.431]	[0.002]	[0.927]	[0.066]
Idle	-0.087*	-0.063	-0.118**	-0.110**	-0.063	-0.117	-0.072
	(0.041)	(0.059)	(0.060)	(0.049)	(0.049)	(0.076)	(0.050)
	[0.034]	[0.286]	[0.048]	[0.026]	[0.194]	[0.125]	[0.153]
Crime	-0.003	-0.007	-0.001	0.131**	-0.134**	0.006	-0.008
	(0.041)	(0.055)	(0.062)	(0.055)	(0.047)	(0.073)	(0.049)
	[0.941]	[0.904]	[0.992]	[0.017]	[0.004]	[0.935]	[0.864]
Teen parenthood	-0.008	-0.043	0.022	-0.099*	0.084	-0.016	-0.003
	(0.040)	(0.056)	(0.057)	(0.051)	(0.049)	(0.070)	(0.047)
	[0.841]	[0.448]	[0.705]	[0.052]	[0.086]	[0.821]	[0.947]
Poor health	-0.068**	-0.038	-0.100**	-0.061	-0.074*	-0.087*	-0.059
	(0.029)	(0.037)	(0.047)	(0.033)	(0.036)	(0.051)	(0.036)
	[0.019]	[0.313]	[0.031]	[0.068]	[0.039]	[0.090]	[0.102]

Table 6: PANEL A - THE EFFECT OF HEAD START OVERALL AND BY SUBGROUP

Panel A: Overall	**5–6 (Test)**	**7–10 (Test)**	**11–14 (Test)**	**5–14 (Test)**	**5–14 (Non-test)**	**19+ (Long term)**
Head Start	0.146	0.128	0.050	0.096	0.208	0.237
(SE)	(0.089)	(0.064)	(0.065)	(0.060)	(0.090)	(0.081)
p-value	0.101	0.045	0.440	0.111	0.020	0.003
Other preschool	-0.035	0.034	-0.025	-0.003	0.095	0.072
(SE)	(0.087)	(0.066)	(0.072)	(0.064)	(0.098)	(0.077)
p-value	0.688	0.610	0.725	0.968	0.330	0.348
p (HS = preschool)	0.061	0.228	0.342	0.189	0.321	0.102

2009]). The reestimated Panel A of Table 4 is shown below. We observe that individual changes are minimal (the vast majority are within the last significant digit), and the overall trend remains consistent with the original results.

Table 7: RESULTS USING DIFFERENT ELIGIBILITY CRITERIA

Panel A: Overall	**5–6 (Test)**	**7–10 (Test)**	**11–14 (Test)**	**5–14 (Test)**	**5–14 (Non-test)**	**19+ (Long term)**
Head Start	0.143	0.127	0.048	0.094	0.214	0.242
(SE)	(0.090)	(0.064)	(0.066)	(0.061)	(0.091)	(0.082)
p-value	0.113	0.047	0.468	0.121	0.018	0.003
Other preschool	-0.037	0.037	-0.025	-0.002	0.094	0.074
(SE)	(0.087)	(0.067)	(0.072)	(0.064)	(0.098)	(0.078)
p-value	0.672	0.581	0.728	0.980	0.338	0.343
p (HS = preschool)	0.066	0.251	0.363	0.207	0.296	0.099

4.2 Covariate Balance Check

Since [Deming, 2009] does not consider unmeasured confounding in the model, we have to pay extra attention to the ignorability assumption. Under the ignorability assumption, we have $T \perp X|e(X)$ where $e(x) = P(T = 1|X = x)$

is the propensity score. Therefore, we can check the ignorability assumption by examining whether the covariate distributions are the same between the treatment and control groups at different strata of the propensity score. Formally, we fit a logistic regression model on all the features and indicators to determine whether the feature is missing (and thus interpolated) to obtain the estimated propensity score. We then stratify the samples into 5 strata based on the quintiles of the estimated propensity score and compare the covariate distributions (including the missing indicators) between the treatment and control groups within each stratum. We consider the Head Start and other pre-school program participation groups separately against the no pre-school program participation group. We use the multivariate extension of the two-sample Kolmogorov-Smirnov (KS) test [Naaman, 2021] and the Bonferroni corrected component-wise univariate two-sample KS test (due to the issue of multiple testing, the Bonferroni corrected p-value is the smallest dimension-wise p-value multiplied by the number of dimensions, capped at 1). The resulting p-values are shown in Table 8. Note that a high p-value indicates that the test fails to reject the null hypothesis that the two distributions are the same.

We observe that the p-values are either 1 or relatively close to 0. We believe that the reason why the p-values of all multivariate KS tests and most Bonferroni-corrected univariate KS tests are 1 is that the covariate dimension is high. Including the missing indicators, the total dimension is 55 (gender is recorded for all children, so the missing indicator is not included). The relatively small p-values for three Bonferroni corrected univariate KS tests are misleading because in all three cases, at most 2 of the 55 univariate KS tests (run dimension-wise) have extremely low p-values, whereas all the other dimensions have high p-values (at least higher than 0.1 and about half are close to 1). We visualize the overlaid treatment and control group's covariate distributions over the first two principal components in Figure 1. We observe that the covariate distributions are very similar between the treatment and control groups. Thus, we conclude that we do not find evidence that the data violates the ignorability assumption.

Table 8: RESULTS FOR COVARIATE BALANCE CHECK

Strata	HStoNone, KS P-Value	HStoNone, Bonferroni KS P-Value	PretoNone, KS P-Value	PretoNone, Bonferroni KS P-Value
0	1.0	0.1676	1.0	1.0
1	1.0	0.0948	1.0	1.0
2	1.0	1.0	1.0	0.1396
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0

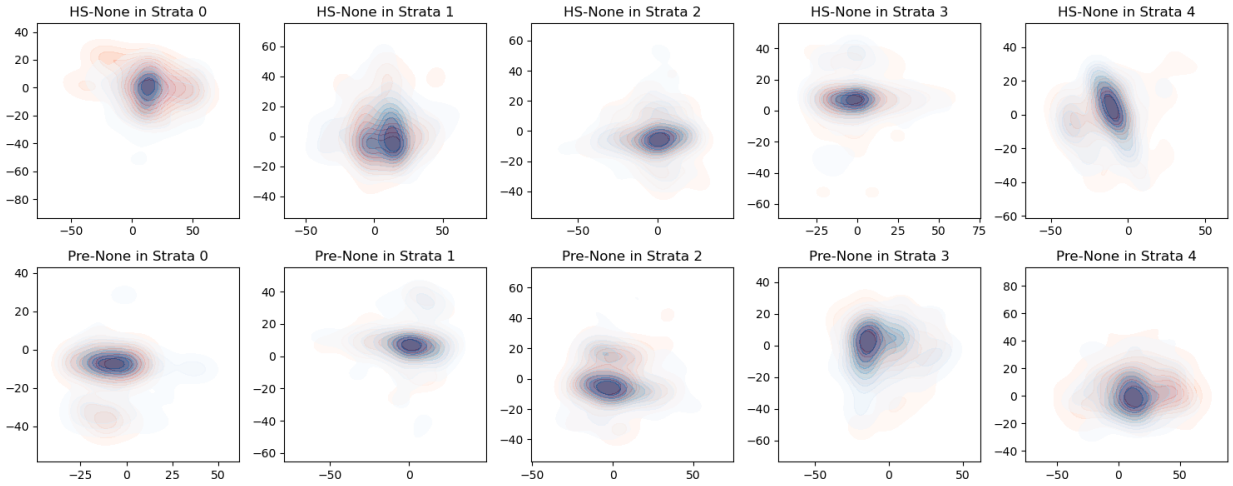

Figure 1: Kernel Density Estimates of Covariate Distributions, Red is Treatment, Blue is Control

5 Reanalysis

In Section 4, we examined the ignorability assumption. In this section, we are motivated by examining the correct modeling assumption 1. Admittedly, the model [Deming, 2009] uses is flexible since it incorporates family fixed effects as well as 55 covariates; however, it still assumes a constant treatment effect. In this section, we explore alternative modeling assumptions and use OLS with interaction, Inverse Probability Weighting (IPW), and Doubly Robust (DR)

estimation methods to estimate the treatment effect. As in the previous section, we only replicate Panel A of Table 4 for brevity.

5.1 OLS with Interaction

The original OLS model 1 ignores treatment effect heterogeneity induced by the covariates. We extend the model by including first-order interaction terms between treatment and each covariate. The new model is given by

$$Y_{ij} = \alpha + \beta_1 HS_{ij} + \beta_2 PRE_{ij} + \delta X_{ij} + \delta_1 HS_{ij} X_{ij} + \delta_2 PRE_{ij} X_{ij} + \gamma_j + \epsilon_i$$

Unlike the original model, the new model requires us to center the covariates. After centering, β_1, β_2 gives the average treatment effect for Head Start and other pre-school program participation, although the consistency of the two estimators still requires the ignorability assumption and the correctness of the new model. Similar to the original model, we report the t-test p-values for $\beta_1 = 0$ and $\beta_2 = 0$, and report the F-test p-values for $\beta_1 = \beta_2$. The results are shown in Table 9. We observe that the individual point and standard error estimates are highly consistent with the original Table 6 (differences are all within 0.1), and the overall trend remains consistent. This shows that including interaction terms does not change the results significantly, and the original model is robust to this extension.

Table 9: REANALYSIS USING OLS WITH INTERACTION TERMS

Panel A: Overall	**5–6 (Test)**	**7–10 (Test)**	**11–14 (Test)**	**5–14 (Test)**	**5–14 (Non-test)**	**19+ (Long term)**
Head Start	0.147	0.146	0.073	0.114	0.196	0.276
(SE)	(0.097)	(0.070)	(0.073)	(0.068)	(0.106)	(0.088)
p-value	0.130	0.038	0.313	0.095	0.063	0.002
Other preschool	0.011	0.084	0.014	0.042	0.085	0.050
(SE)	(0.092)	(0.071)	(0.073)	(0.068)	(0.098)	(0.080)
p-value	0.906	0.239	0.844	0.533	0.389	0.532
p (HS = preschool)	0.207	0.479	0.499	0.396	0.380	0.031

5.2 IPW

Both the original OLS model and the OLS with interaction model assume a linear, parametric relationship between the covariates, treatment, and outcome. In this subsection, we are interested in exploring a non-parametric identification strategy. We use the Hajek estimator, which is a translation-invariant extension of the IPW estimator. Concretely, we fit a propensity score estimator $\hat{e}$ through logistic regression, and produce

$$\hat{\tau}_{hajek} = \frac{\sum_i \frac{Y_i T_i}{\hat{e}(X_i)}}{\sum_i \frac{T_i}{\hat{e}(X_i)}} - \frac{\sum_i \frac{Y_i (1-T_i)}{1-\hat{e}(X_i)}}{\sum_i \frac{1-T_i}{1-\hat{e}(X_i)}}$$

The Hajek estimator requires the ignorability assumption and the correctness of the propensity score model (which we verify in Section 4). We separately consider the Head Start and other pre-school program participation groups against the no pre-school program participation group and use a non-parametric bootstrap to estimate the standard error and the p-values. The results are shown in Table 10. We observe that the point estimates are usually lower than the original Table 6, the head start effect estimates are less significant, the other pre-school effect estimates remain relatively insignificant, and the difference between head start and other pre-school program's effects are less significant. Nevertheless, under the Hajek estimator, we still observe that pre-school programs' effect on test scores fades out by age, non-test and long-term effects of pre-school programs persist, and Head Start has a greater effect than other pre-school programs. These trends are consistent with the original table and constitute the main results of [Deming, 2009].

5.3 Doubly Robust

Having explored linear, parametric models with OLS and non-parametric models with IPW, we are interested in exploring the Doubly Robust (DR) estimator, which assumes ignorability and either the correct model for the propensity score or the correct model for the outcome. Concretely, we fit a propensity score estimator $\hat{e}$ through logistic regression and use the vanilla OLS model with fixed effects (but without interaction terms). We estimate the mean potential outcomes $\mathbb{E}[Y(1)|X]$ with $\hat{\mu}_1(X)$, the value of the OLS model evaluated at covariates X and $T = 1$. Similarly, we estimate $\mathbb{E}[Y(0)|X]$ with $\hat{\mu}_0(X)$, the value of the OLS model evaluated at covariates X and $T = 0$. We produce the

Table 10: REANALYSIS USING THE HAJEK ESTIMATOR

Panel A: Overall	5–6 (Test)	7–10 (Test)	11–14 (Test)	5–14 (Test)	5–14 (Non-test)	19+ (Long term)
Head Start	0.127	0.047	-0.046	0.021	0.078	0.122
(SE)	(0.083)	(0.068)	(0.067)	(0.059)	(0.076)	(0.078)
p-value	0.053	0.237	0.753	0.349	0.151	0.059
Other preschool	0.025	-0.010	-0.074	-0.030	0.014	0.075
(SE)	(0.092)	(0.068)	(0.076)	(0.062)	(0.071)	(0.072)
p-value	0.386	0.569	0.835	0.681	0.414	0.146
p (HS = preschool)	0.299	0.460	0.745	0.447	0.438	0.580

DR estimator

$$\hat{\tau}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i\left(Y_i - \hat{\mu}_1(X_i)\right)}{\hat{e}(X_i)} + \hat{\mu}_1(X_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-T_i)\left(Y_i - \hat{\mu}_0(X_i)\right)}{1-\hat{e}(X_i)} + \hat{\mu}_0(X_i)\right)$$

As for the IPW estimator, we separately consider the Head Start and other pre-school program participation groups against the no pre-school program participation group and use a non-parametric bootstrap to estimate the standard error and the p-values. The results are shown in Table 11. We observe that for test scores, the estimated effect for both Head Start and other pre-school programs ranges in a wider range compared to the original Table 6, while preserving the trend of effect fading out. The point estimates for non-test and long-term effects are similar to the original. The p-values are generally lower than or similar to the original p-values, with the exception of the test for Head Start and other pre-school program's effect differences.

Table 11: REANALYSIS USING THE DOUBLY ROBUST ESTIMATOR

Panel A: Overall	5–6 (Test)	7–10 (Test)	11–14 (Test)	5–14 (Test)	5–14 (Non-test)	19+ (Long term)
Head Start	0.307	0.141	0.072	0.109	0.252	0.263
(SE)	(0.185)	(0.093)	(0.094)	(0.078)	(0.120)	(0.106)
p-value	0.038	0.060	0.221	0.085	0.020	0.005
Other preschool	-0.020	0.006	0.019	0.005	0.089	0.064
(SE)	(0.167)	(0.093)	(0.093)	(0.077)	(0.120)	(0.094)
p-value	0.538	0.479	0.419	0.483	0.225	0.239
p (HS = preschool)	0.109	0.193	0.628	0.262	0.277	0.122

5.4 Conclusion

All of the three alternative estimators give reanalysis results similar to the original model, preserving the overall trend and thus the main conclusion of [Deming, 2009]. Thus, we conclude that the main conclusion of the paper is robust.

Since the OLS model 6 gives estimates that are more similar to the Doubly Robust estimator 11 compared to the IPW estimator 10, we hypothesize that the OLS with fixed effects model for the potential outcomes is more correct than the logistic regression model for the propensity score. Since family fixed effects give the OLS model great flexibility, this observation does not seem surprising.

6 Appendix

We provide key summarization statistics of the outcome and the covariates in Table 12. The first three rows are the test scores, non-test scores, and long-term outcomes (all standardized) aggregated over various individual outcomes. We observe that the mean of all three outcomes in the other pre-school program participation group is the highest, followed by no pre-school program participation group, and then the Head Start group. This implies that the Head Start participants are, on average, worse performing than their peers. However, after adjusting for confounding, the Head Start group has a consistently positive effect on all three outcomes compared to other pre-school program participation and no pre-school program participation.

The next 28 rows are the covariates used throughout the analysis. We notice that there are a large amount of missing values, and we impute them with the mean of the non-missing values conditioned on the same race and gender. In general, head start participants come from less advantaged families with lower incomes and worse health conditions.

Table 12: Summary Statistics of the Outcomes and Covariates

	Mean	Median	Std Dev	Head Start Mean	Other Preschool Mean	No Preschool Mean	Total Sample Size
Test Scores (5-14, standardized)	-0.000	-0.047	1.000	-0.210	0.228	-0.044	1455
Non-Test Outcomes (5-14, standardized)	0.000	0.620	1.000	-0.059	0.098	-0.049	1455
Long Term Outcomes (18+, standardized)	-0.000	0.162	1.000	-0.118	0.212	-0.116	1455
male	0.491	0.000	0.500	0.494	0.489	0.490	1455
reside_with_mother_0to3	0.689	1.000	0.463	0.689	0.708	0.668	1384
prexisting_health_conditions	0.042	0.000	0.200	0.045	0.038	0.043	1384
very_low_birth_weight	0.017	0.000	0.129	0.018	0.010	0.023	1424
logBW	4.711	4.745	0.253	4.699	4.730	4.702	1424
log_income_0to3	10.047	10.039	0.728	9.814	10.302	10.006	1386
log_income_at_3	9.997	10.024	0.902	9.708	10.274	9.988	1160
first_born	0.423	0.000	0.494	0.434	0.420	0.417	1455
PPVTat3	20.724	17.000	12.941	16.129	23.758	20.127	261
HOME_Pct_0to3	40.470	35.750	27.369	32.503	47.832	37.474	530
Moth_HrsWorked_BefBirth	26.165	27.769	11.971	23.295	27.697	26.139	475
Moth_HrsWorked_Avg_0to3	32.221	35.467	10.766	32.547	32.154	32.021	816
Moth_HrsWorked_0to1	32.076	36.833	11.969	31.233	32.375	32.420	478
Father_HH_0to3	0.635	1.000	0.447	0.498	0.747	0.623	887
GMom_0to3	0.210	0.000	0.330	0.253	0.165	0.214	1387
MomCare	0.643	1.000	0.413	0.676	0.577	0.677	1448
RelCare	0.197	0.000	0.337	0.174	0.227	0.188	1448
NonRelCare	0.161	0.000	0.300	0.150	0.196	0.135	1448
Moth_Smoke_BefBirth	0.359	0.000	0.480	0.340	0.349	0.388	1384
Alc_BefBirth	0.235	0.000	0.424	0.292	0.196	0.235	502
Breastfed	0.344	0.000	0.475	0.238	0.445	0.344	1436
Doctor_0to3	0.382	0.000	0.486	0.479	0.316	0.375	537
Dentist_0to3	0.292	0.000	0.455	0.372	0.235	0.291	504
Moth_WeightChange	30.019	28.000	15.311	28.455	31.476	30.020	1338
Illness_1stYr	0.528	1.000	0.499	0.530	0.531	0.524	1370
Premature	0.210	0.000	0.407	0.214	0.205	0.210	1372
Insurance_0to3	0.525	0.500	0.482	0.353	0.687	0.456	538
Medicaid_0to3	0.312	0.000	0.448	0.506	0.136	0.379	538

References

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

Janet Currie and Duncan Thomas. Does head start make a difference? *The American Economic Review*, 85(3):341–364, 1995.

Michael Naaman. On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters*, 2021.