ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-27

Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts

Remy J.-C. Pages
University of California Irvine

Dylan J. Lukes
Harvard University

Drew H. Bailey
University of California Irvine

Greg J. Duncan
University of California Irvine

Using an additional decade of CNLSY data, this study replicated and extended Deming's (2009) evaluation of Head Start's life-cycle skill formation impacts in three ways. Extending the measurement interval for Deming's adulthood outcomes, we found no statistically significant impacts on earnings and mixed evidence of impacts on other adult outcomes. Applying Deming's sibling comparison framework to more recent birth cohorts born to CNLSY mothers revealed mostly negative Head Start impacts. Combining all cohorts shows generally null impacts on school-age and early adulthood outcomes.

VERSION: May 2019

Suggested citation: Pages, J.C., Lukes, D.J., Bailey, D.H., & Duncan, G.J. (2019). Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts (EdWorkingPaper No.19-27). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-27

Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts

By REMY J.-C. PAGES, DYLAN J. LUKES, DREW H. BAILEY AND GREG J. DUNCAN*

April 2, 2019

Using an additional decade of CNLSY data, this study replicated and extended Deming's (2009) evaluation of Head Start's life-cycle skill formation impacts in three ways. Extending the measurement interval for Deming's adulthood outcomes, we found no statistically significant impacts on earnings and mixed evidence of impacts on other adult outcomes. Applying Deming's sibling comparison framework to more recent birth cohorts born to CNLSY mothers revealed mostly negative Head Start impacts. Combining all cohorts shows generally null impacts on school-age and early adulthood outcomes.

Understanding the causal effects of early childhood programs implemented at scale on long-term adult outcomes is challenging. But because early childhood is considered by many economists to be a key launching period for life-long human capital accumulation (e.g., Cunha et al. 2006; Chetty et al. 2011; Currie and Almond 2011; Heckman and Mosso 2014; Hoynes, Schanzenbach and Almond

* Remy J.-C. Pages and Dylan J. Lukes are co-equal first authors. Before collaborating, both individuals were separately and contemporaneously working on individual papers which substantially overlapped. After careful consideration all parties agreed to collaborate and produce a single unified paper. We thank Felipe Barrera-Osorio, Damon Clark, David Deming, Chloe Gibbs, Josh Goodman, Eric Taylor for feedback on prior drafts and presentations of this project. Contact information for each of these authors can be viewed below: School of Education, University of California, Irvine, 3200 Education, Irvine, CA 92697 (e-mail: rpages@uci.edu) and School of Education, Harvard University, 13 Appian Way, Cambridge, MA 02138 (email: dylanlukes@g.harvard.edu).

2016), considerable attention has been devoted to research attempting to estimate the short and longer-run impacts of early education programs (Duncan and Magnuson 2013).

Because Head Start is the U.S.'s oldest and largest early childhood education program to be offered at scale, it is hardly surprising that a great deal of the research in this area has been devoted to it. Studies of the longer-run impacts of Head Start attendance have shown generally positive although sometimes mixed results (Garces, Thomas and Currie 2002; Ludwig and Miller 2007; Deming 2009; Carneiro and Ginja 2014; Bauer and Schanzenbach 2016; Thompson 2018). In a recent study analyzing data from the Panel Study of Income Dynamics survey, Johnson and Jackson (2018) suggested that some of these inconsistencies can be attributed to complementarities between Head Start attendance and subsequent K-12 spending.[1]

Using data from the National Longitudinal Survey of Youth 1979 Children and Young Adults (thereafter, CNSLY), and building on the approach of Currie, Thomas (1995; also, Garces et al. 2002), Deming's (2009) study drew sibling data from 1970-1986 birth cohorts and measured outcomes through the 2004 interviewing wave. Deming compared school-age and young adulthood outcomes for children who attended Head Start and their siblings who either attended other non-Head Start preschools or did not attend any preschool education program. He estimated that, compared with siblings who did not attend any preschool, children who attended Head Start had on average 8.5 percentage points higher rates of high school graduation and 0.23 standard deviations (sd) higher scores on an index of adult outcomes.[2] Deming's study is noteworthy both for its sibling comparison

[1] For example, Johnson and Jackson calculated that for a child attending Head Start, a 10 percent increase in K-12 spending boosted educational attainment by 0.4 years, earnings by 20.6 percent; and reduced the probability of being incarcerated by 8 percentage points.

[2] This composite measure included high school graduation, college attendance, teen-age parenthood, idleness (i.e. neither working nor attending school), involvement with the justice system and poor health status.

design, which controls for some unmeasured time-invariant factors of the family environment, and because of its use of a reasonably large and relatively recent national longitudinal sample followed through childhood into early adulthood.

Evidence of Head Start's lasting positive impact is not limited to Deming (2009). Over the past several decades, using a variety of empirical methods including family fixed effects, regression discontinuity and difference-and-difference, research has accumulated mounting evidence of Head Start's ability to improve adolescent and longer-term outcomes (Currie and Thomas 1995, 1999; Garces et al. 2002; Ludwig and Miller 2007; Carneiro and Ginja 2014; Bauer and Schanzenbach 2016; Bailey, Sun & Timpe 2018; Barr and Gibbs 2018; Miller, Shenhav and Grosz 2018; Thompson 2018). A detailed synthesis of these studies, including birth cohorts analyzed, empirical design and findings can be found in the Appendix. Most consistent across this set of research is Head Start's positive impact on educational attainment, health outcomes and reduced criminal activity. Many of these effects are larger and more robust for samples composed of mothers with less than a high school education, earlier birth cohorts, and males. These studies are predominantly composed of 1970s and 1980s birth cohorts. However, notable exceptions include Carneiro and Ginja (2014), Bauer and Schanzenbach (2016) and Barr and Gibbs (2018), which also include birth cohorts from the early 1990s.[3]

The present work builds upon this rich set of existing literature by expanding Deming's (2009) evaluation of Head Start's longer-run impacts. By appending 10 additional years to the original 1970-1986 birth cohorts analyzed in Deming (2009), we were able to estimate impacts on outcomes measured later in adulthood and not considered in Deming (2009): educational attainment, college graduation, and earnings. Second, the additional data provided us with an opportunity to apply

[3] Carneiro and Ginja (2014) RD analysis sample includes 1977-1996 birth cohorts, Bauer and Schanzenbach (2016) FFE analysis sample includes 1970-1990 birth cohorts, and Barr and Gibbs (2018) FFE analysis sample used in a supplementary exercise in their appendix includes 1970-1992 birth cohorts. NLSY and CNLSY data used in each. The FFE analysis sample used in this paper includes 1970-1996 birth cohorts from CNLSY.

Deming's methods to 10 additional birth cohorts in the NLSY to address whether his results generalized to cohorts born to older mothers and into somewhat different historical conditions. Third, we estimated impacts on both school-age and adulthood outcomes for a sample combining all possible cohorts to provide estimates based on the broadest population base.

We found that extending the measurement period for Deming's cohorts and early-adult outcomes decreased the estimated impact on the adulthood index of Head Start attendance relative to not attending any preschool program from 0.23 sd to 0.17 sd (standard error (se) = 0.07). Of the longer-run outcomes we were able to consider, the largest impact of attending Head Start was on years of completed schooling (0.30 years; se = 0.15). However, impacts on college graduation and earnings (computed as the log of 1994-2014 averaged earnings, adjusted for age and survey-year) were small and not statistically significant.

For the cohorts born after Deming's cohorts, Head Start impacts were mostly null and sometimes negative. In fact, positive impacts on the adulthood summary index generated by Deming's cohorts were matched by nearly symmetric negative impacts for the complement cohorts (e.g., -0.15 sd, se = 0.07). For the final sample that combined the two sets of cohorts, the point estimate of Head Start's impact on the adult summary index was close to zero and not statistically significant. We also observed and documented differences in baseline human capital between cohorts. Although not explicitly tested, there are hints in more recent research that Head Start's impact on long-run outcomes may have diminished across time, favoring participating children from earlier 1970s and 1980s birth cohorts compared with later 1990s birth cohorts (Carneiro and Ginja 2014, Bauer and Schanzenbach 2016, Barr and Gibbs 2018).[4]

[4] Carneiro and Ginja (2014) indicate in a robustness test that most of the effects for males age 12-13 in their overall sample are driven by the earlier 1980s birth cohorts. In addition, Barr and Gibbs (2018) appendix FFE analysis find no significant impact of Head Start on high school graduation, some college, crime, teen parenthood or their index of adulthood outcomes. However, they do find positive and significant impacts for males on high school graduation, crime and their index

Thus, although in the past several years there has been a reemergence of research that revisits the long-run impacts of Head Start, this study differentiates itself in several key ways. First, we use a well-established method that hitherto has estimated positive long-run outcomes of Head Start to estimate predominantly negative or null long-run outcomes of the program. Second, to our knowledge, this is the first paper to estimate the impact of Head Start for the most recent set of CNLSY birth cohorts and explicitly analyze how program effects have changed for them, despite major changes to the program and social context of Head Start-eligible children during this period.[5]

A brief review of the literature on Head Start's short and medium-term evaluations is presented in the online appendix S1. Below, section I presents some relevant background on Head Start. Section II describes the data. Section III reintroduces Deming's sibling comparison model, as well as the pre-treatment covariates selected to probe and mitigate selection bias. Section IV organizes the results by outcomes and cohorts. Section V summarizes the findings and concludes.

I. Head Start Program Background

Part of the Johnson administration's Great Society policies, Head Start was launched in 1965 to provide educational and health-related services to children living in low-income families. As of 2017, about 900,000 children were enrolled in Head Start, 97 percent of whom were between the ages of 3 and 5, at a cost of around $9 billion in federal funding (U.S. Department of Health and Human

of adulthood outcomes. In both cases, each overall sample included birth cohorts from 1970s through early to mid-1990s. In contrast, Bauer and Schanzenbach (2016) find positive impacts of Head Start on high school graduation, some higher education, post-secondary completion, self-control index, self-esteem index and positive parenting index. These results more closely follow Deming (2009) but only include birth cohorts up to 1990.

[5] As mentioned above, Carneiro and Ginja (2014), Bauer and Schanzenbach (2016), and Barr and Gibbs (2018) all have analyses of Head Start impacts that include 1990s birth cohorts. However, these studies do not systematically estimate Head Start impacts by birth cohorts across time. Instead primary results are for analyses based on overall samples which include both more recent birth cohorts from the late 1980s and early 1990s, and older birth cohorts from 1970s and early to mid-1980s.

Services 2018). Enrollment and funding have varied greatly since Head Start's 1965 inception. Participation grew until the early 1980s, plateaued through the early 1990s and then grew again when funded enrollments almost doubled (i.e. around 500,000 in 1990 to 900,000 in 2000). Appropriations (in 2018 USD) grew from about $3 billion in 1990 to $9 billion in 2000. After 2000, both enrollment and inflation-adjusted funding remained steady (DHHS 2018).

Between the 1989-1990 (which are typical Head Start attendance years for Deming's cohorts) and 1996-1997 (which are typical attendance years for our complement cohorts) school years, enrollment increased by about 60 percent. However, the proportion of teachers or assistant teachers with at least a Child Development Associate credential increased very little – by about 5 percentage points over this period (DHHS 2018). More generally, the 1990s and 2000s were a time of rapid increases in preschool enrollment, including Head Start, but also state-run pre-kindergarten programs (Duncan and Magnuson, 2013).

II. Data

Figure 1 provides an overview of birth years and years in which childhood and adult outcomes are measured for the two sets of cohorts that form our analytic samples. Deming's cohorts were born between 1970 and 1986 and attended Head Start no later than 1990. Moreover, Deming's sibling fixed effects analyses were estimated for a sample of siblings discordant on Head Start attendance and who enrolled in Head Start no later than 1990. His sample eligibility rules are: 1) at least two children aged 4 or older by 1990 within the same family; and 2) at least one pair of siblings in a family had to be discordant across Head Start, other preschool, or neither statuses. The median age of individuals in Deming's analytic sample was 23 years (21 and 25 years for first and third quartile, respectively) by 2004, the most recent CNLSY survey round year available for his study.

[Insert Figure 1 here]

For our complement and combined cohorts, Deming's sample restrictions were moved forward by ten years: samples were restricted to siblings who were at least 4 by 2000 (i.e., at least 19 by 2014). Sample restrictions produced sample sizes of N = 1,251 for Deming's cohorts, N = 2,144 for our complement cohorts and N = 3,768 for our combined cohorts.[6] It is important to note that the fertility dynamics of the CNLSY (all children were born to women who were between ages 14 and 22 in 1979) leads children in our complement cohort to be born to older mothers than is the case for children in Deming's cohort.

Because we wanted to both estimate Head Start impacts on educational attainment, college graduation and earnings, and assess the impacts' robustness on the multiple outcomes adulthood index, we both replicated and extended Deming's analysis of this cohorts up to 2014, the latest CNLSY survey round year available to us.

In Table 1, household characteristics are presented by cohort and preschool status (Head Start vs. the counterfactual of no preschool); permanent income; maternal education and cognitive test score[7]; and grandmothers' highest grade completed. Across these variables and for all three cohorts, there was a clear negative selection into Head Start for samples of siblings under rule 1 only—a less restricted sample, more representative of the CNLSY sample—and samples with rule 2 added (i.e., the fixed effects subsamples). Discrepancies between the two samples were small,

[6] A third restriction was applied to the complement cohort: siblings were considered for eligibility up to 2000 excluding those already part of Deming's cohort (i.e., selected under rule 1 and 2); it is in that sense that this new cohort is the complement of Deming's cohort. Of all siblings composing the complement cohort, 78 percent had reached four years of age post 1990 (75 percent, for Head Start participants). As in Deming (2009) and for all cohorts, the original NLSY79 oversample of low-income whites was excluded. The sum of the first two sample sizes is smaller than the size of the combined cohorts sample because, under the latter configuration, there are more opportunities for siblings to fit sample eligibility criteria (described further below).

[7] NLSY79 derived, from the Armed Services Vocational Aptitude Battery of tests, the Armed Forced Qualification Test (AFQT) comprising items in arithmetic reasoning; mathematics knowledge; word knowledge; and paragraph comprehension.

suggesting that the demographic characteristics of the fixed effect subsamples were similar to the less restricted, larger samples.[8]

[Insert Table 1 here]

As shown in the column 'Difference HS-None'—reporting mean differences in standard deviation units for Deming's cohorts, the complement cohorts and the combined cohorts, respectively—selection into Head Start was similarly associated with socioeconomic disadvantage for Deming's cohort as well as for the complement cohort. For example, Head Start participants had a 0.44 sd lower permanent income and a 0.59 sd lower maternal AFQT than children not attending any preschool. Overall, Head Start children came from relatively more disadvantaged households.[9] As noted by Deming (2009), because his cohort of Head Start participants had been born to younger mothers (their median age was 20), they might have benefited more from the program (which, in addition to early education, includes services for parents). In contrast, for the complement cohort, mothers were older (median age was 28), and household characteristics more favorable on all of the dimensions included in Table 1.

B. Outcomes

As part of our replication of Deming's (2009) study, we developed comparable outcome variables for each cohort. Taken from age 5 to 14, three cognitive test measures (the Peabody Picture Vocabulary Test; the Peabody Individual

[8] Deming (2009) presented these characteristics over three preschool status (Head Start; other preschool; no preschool) and by racial/ethnic subgroups. To keep Table 1 manageable, only overall means for Head Start and no preschool status (the counterfactual) are displayed (for details with other preschool status included, and by racial/ethnic subgroups, see online appendix Tables S2-S4). There was a reduction in sample sizes when restricting on families who made difference choices of preschool status for at least 2 siblings (rule 2 in section III): by 66 percent for Deming's cohort, by 41 percent for the complement cohort; and by 45 percent for the combined cohorts. However, variation on selected household characteristics appeared to be very similar across both type of samples, and across all cohorts (Table 1).

[9] Differences were even more pronounced when comparing between Head Start and the other "Preschool" status (see online appendix Table S2).

Achievement math and reading recognition subtests) were averaged to form a ‘Cognitive tests index’. Over the same age period, the CNLSY mother-reported Behavior Problems Index (BPI) was also included. From age 7 to 14, a ‘Nontests index’ was built from two indicators: grade retention and learning disability diagnosis.

To reduce the risks of multiple-inference inflated Type I errors, Deming developed a multiple outcomes adulthood index, which we also replicated. It consists of an equally-weighted standardized composite of high school graduation; teen parenthood; some college attended; idleness; involvement with the justice system; poor health status. All outcomes were measured up to the CNLSY 2014 survey-round. For Deming’s cohort, median age was 33 (all individuals being at least 28). For the complement cohort, the median age was 25; with 75 percent of that cohort 23 or older. For the combined cohorts, median age was 29; 75 percent of whom were 25 or older.

The longer time series of NLSY data enabled us (but not Deming) to estimate Head Start effects for Deming’s cohort on completed years of schooling and college graduation[10], and on earnings. The earnings composite for each sample member was obtained by first pooling all person-year earnings observations (in 2014 dollars) and then regressing them on dummy-variable indicators for birth cohort and calendar year to purge earnings of birth cohort and measurement year effects.[11] From the coefficients in this regression, we generated a set of person-year earnings residuals for all individuals in the analysis sample. We then averaged these earnings

[10] Both variables were derived from CNLSY cross-round item asking respondents which highest grade they had completed at the date of the latest survey round interview. Responses were recoded as equivalent years of completed schooling (e.g., if respondent answered “high school graduate” it was recoded 12; ‘completed an associated degree’ was recoded 14; etc.).

[11] We pooled older (35+ age) respondents together (N = 589) to ensure a sufficient sample size for these cohorts before regressing for adjustment. For all other birth cohorts, sample sizes were of at least 300 observations. The arguably arbitrary 25-year threshold was chosen such that *a priori* valid inputs would be available.

residuals for each individual, added them to the grand mean earnings in the sample, and took the natural logarithm of this earnings average.

III. Empirical Strategy and Selection Bias

As noted, families selecting into Head Start were relatively more disadvantaged on a series of selected household characteristics. Consequently, Head Start estimates relative to other preschool status based on cross-family variation may be negatively biased. A family fixed-effects design mitigates some of these biases by separating the potentially confounding influence of family environment variance shared among siblings from estimations of interest. This was the empirical strategy undertaken in Deming (2009), which we reproduced in the present study and formalized in the same fashion:

$$(1) \qquad Y_{ij} = \alpha + \beta_1 HS_{ij} + \beta_2 PRE_{ij} + \delta \mathbf{X}_{ij} + \gamma_j + \varepsilon_i.$$

In this model, i and j respectively index individuals and families. Thus, HS_{ij} (PRE_{ij}) stands as an indicator for participation in Head Start (Preschool) where β_1 (β_2) denotes Head Start (Preschool) impact estimates on outcome Y_{ij}, for some sibling i within family j, relative to a sibling (within family j) attending neither. $\mathbf{X}_{ij}$ represents the vector of 'pre-treatment' family covariates pertaining to sibling i within family j; family j fixed-effect is captured by γ_j, while ε_i represents sibling i's residual.

Within-family comparisons remove the effects of time-invariant family characteristics on sibling's outcomes. There remains, however, a strong possibility of within-family selection bias since siblings within the selected families of our cohorts differed in their preschool status allocation. Reasons for such choices were not recorded in the CNLSY. To mitigate such potential for bias, Deming (2009) opted for a series of sibling-specific family-level covariates—the ones represented by the vector **X** of equation (1)—measured before or at Head Start preschool program eligibility age (3 years old). We examined these covariates for the complement and combined cohorts and tested whether siblings within a given family systematically differed on these covariates.[12] Within our family fixed-effects framework, each covariate was thus regressed on siblings' preschool status, either 'Head Start' or 'Preschool'. A statistically significant and substantial variation from 'No preschool' (the reference status) would then signal a potential selection bias regarding the relation between that pre-treatment characteristic and the regressed on preschool status. These regression estimates were reported in the online appendix, Table S5.

Focusing first on the complement cohort, siblings attending Head Start were on average older by one year, and by almost two years for siblings attending other preschools, than their counterparts not attending any preschool program. This was consistent with the probability for a first-born sibling to be enrolled in preschool being greater by 10 percentage points for Head Start enrollees, and by 28 percentage points for other preschool participants. Both groups were 6 to 8 percentage points less likely to receive maternal care from birth to age three, and so more likely to receive care from a non-relative (5 and 6 percentage points, respectively). Attrition was low for both the complement and the combined cohorts, averaging about 4 and 3 percent, respectively. Moreover, both Head Start and other preschool participants were somewhat less likely (by about 3 and 2 percentage points, respectively) to be part of observations lost to attrition.

To characterize selection bias as it pertains to overall disadvantage, a summary index of all pre-treatment covariates was constructed in the same ways as the

[12] As in Deming (2009), when estimating Head Start impacts in the regression models, missing data for these covariates were imputed with corresponding sample mean value. For each, a dichotomous indicator for imputed responses was also included.

multiple outcomes adulthood index described earlier.[13] For all cohorts, as with the covariates, the pre-treatment index was regressed on two preschool indicators, keeping the no preschool status as the reference category. We found that Head Start and other preschool within-family effects on the pre-treatment index were close to zero and never statistically significant. As with Deming (2009) selection bias analysis (which we replicated, see online appendix Tables S6-S7), we could not reject the null hypothesis of equality between preschool statuses.

There was no evidence of parental selection into Head Start (or any other preschool status) within each cohort and household, Figure 2 illustrates what might be one explanation for cross-cohort differences of Head Start impacts on longer-run adulthood outcomes. Complement cohort respective kernel densities of pre-treatment index scores were shifted to the right, i.e., towards more favorable household characteristics for the complement cohort. Complement cohort siblings having attended Head Start later would have then, on average, benefited from more household resources. Compared with Deming's cohort, such a shift might stand as a potential substitute for whatever impact the program would have otherwise yielded.

[Insert Figure 2 here]

IV. Results

Each of the following subsections is organized by cohorts. We first present Head Start impacts on the adulthood summary index, along with its individual composing outcomes. Second, longer-run Head Start impacts on adulthood summary index,

[13] As in Deming (2009), variables composing the pre-treatment covariates index were all first positively oriented with respect to the adulthood summary index. For example, variables like gender (male), age (older), or grand-mother living in household between child's birth and age 3, were negatively correlated with the outcome. Their correlational direction was reversed multiplying their sign by -1. All covariates were then standardized and aggregated into an index, in turn, also standardized (mean = 0; s.d. = 1).

educational attainment, college graduation and earnings are described. Third, estimates for school age outcomes are shown. Robustness checks are presented in the final subsection.

A. *Head Start Impacts on the Adulthood Summary Index*

In Table 2, the family fixed effects model was implemented in steps[14]—with three model specifications; repeated across the complement and the combined cohorts—to gauge the relative directions of biases from observed covariates and unobserved household-level confounders. Model (1) included no family fixed effects but included the pre-treatment covariates from Appendix Table B1, along with household predictors (Table 1)—namely, standardized permanent income; maternal AFQT score; one indicator for maternal high school graduation and one for some college attendance. By contrast, model (2) includes only family fixed effects. Model (3) includes both fixed effects and pre-treatment covariates. Moving from model (1) to model (2), the explained variance (R^2) was larger for all cohorts. Hence, error variance from unobserved variables was smaller than that from the selected observed variables. Moreover, within the R^2 column, including pre-treatment covariates to the fixed effects model—i.e., moving from model (2) to model (3)—added some precision to the estimates (explained variance increasing from 0.639 to 0.686; from 0.711 to 0.728; and from 0.612 to 0.628 for Deming's, complement and combined cohorts, respectively).[15]

[Insert Table 2 here]

[14] Deming (2009; pp. 123-124) used a similar approach in estimating impacts on the school age cognitive tests index.

[15] From similar trends, Deming (2009) concluded (based on Altonji, Elder and Taber 2005 seminal work on the topic) that estimates obtained from model (3) stood as lower bounds for Head Start causal impacts.

From the middle panel, for the complement cohort, Head Start impacts on the adult summary composite were negative at -0.145 sd (significant at the 5 percent level). This value was in clear contrast (*p* = 0.005) with Deming's cohort estimate of 0.166 sd (se = 0.069) on that same multiple outcomes adulthood index. In the bottom panel, for the combined cohorts model (3), no Head Start impacts' estimates were statistically significant; most were negative and approaching zero.[16]

We investigated the 0.228 sd (in Deming 2009 with outcomes measured up to CNLSY 2004 survey-round) to 0.166 sd (when outcome measurement was extended to 2014) decrease in impacts on the multiple outcomes adulthood index. The change was due in part to impacts on the indicator 'Idle' changing sign and ceasing to be statistically significant: by 2004, Deming's cohort Head Start participants were 7 percentage points (se = 0.038) less likely to be 'idle'; this impact had disappeared by 2014 (-3 percentage points; se = 0.041). Thus, with the passage of the additional decade, Head Start participants were not, on average, better positioned to pursue a college degree or to have a job, relative to their siblings not having attended any preschool program.

Figure 3 shows Head Start impacts on all individual outcomes composing the adulthood index.[17] For Deming's cohort, 'Poor health status' stayed favorable by 5 percentage points. (se = 0.032), decreasing slightly from 2004, when Deming's cohort were 7 percentage points (se = 0.026) less likely to self-identify as being of

[16] Throughout (see online appendix, Tables S8-S11), we considered identical demographic subgroups as in Deming (2009). Looking at online appendix Table S8 'Adulthood index' column, for the combined cohorts sample most impacts of Head Start were also close to zero. One exception being for siblings with low maternal AFQT subgroup—one s.d. below NLSY79 AFQT empirical sample average (online appendix Table S8, bottom panel). For these siblings (N = 810), Head Start appeared to have had a marginally significant positive impact on the adulthood summary index (0.108 sd; se = 0.080). This estimate was greater for Deming's cohort (0.384 sd, significant at the 1 percent level). For the complement cohort, impact was much smaller (0.027, se = 0.145), although difference testing between this cohort's estimate and that of Deming's cohort did not fall below the 10 percent level of statistical significance (*p* = 0.147). For the complement cohort, the proportion of siblings with low maternal AFQT background was also smaller (0.21 compared with 0.41 within Deming's cohort) which is consistent with the overall observed favorable shift over household characteristics between the two cohorts (see Table 1 in section III).

[17] For all estimates on individual outcomes, overall and by subgroups, see online appendix tables S9-S10.

poor health. Impacts on 'Some college attended' rose to statistical significance (11 percentage points; se = 0.039).[18] On 'Crime', Head Start participants did not appear to have had more involvement with the justice system than their siblings. Yet, impacts on teenage parenthood shifted unfavorably. Since impacts on the adulthood index are based on score averaging across the composing indicators, Deming's cohort overall decline on this index were captured by the changes just described.

[Insert Figure 3 here]

In contrast to Deming's cohort, Head Start impacts on the complement cohort were mostly negative and larger in absolute value; when positive, they were smaller in absolute value. Head Start's estimated impact on 'Idle' was relatively large, negative (-0.080; se = 0.030) and significant at the 1 percent level. Thus, in the complement cohort, siblings who have attended Head Start were less likely by about 8 percentage points to be employed or enrolled in school (by age 19 or older), compared with their siblings who received home care. Impact on 'Some college attended' went in the opposite (negative) direction for the complement cohort (-0.074 sd; se = 0.039), as well as for 'Crime' (reversed scaled, -0.033; se = 0.031).[19] In sum, the discrepancies between the two cohorts over Head Start impacts on these individual outcomes are aligned with the difference observed earlier over the adulthood summary index (Table 2). Once more, impact estimates for the combined cohorts sample were small and never statistically significant.

[18] Education data were obtained using CNLSY 2014 survey-round *cross-round* variable for respondents' highest grade completed.

[19] For both Deming's and complement cohorts, Head Start impacts on 'Poor health status' were positive (i.e., not self-identifying as being of poor health). This is in line with results found on a range of health outcomes in Carneiro and Ginja Head Start evaluation study (2014). The estimate for the combined cohorts was very small though, with a standard error well balanced across zero (Figure 3; online appendix Table S10).

B. *Head Start Impacts on Longer-Run Outcomes*

Head Start longer-run impacts are displayed in Figure 4 (the complete set of estimates is presented in online appendix Table S11). As mentioned in the introduction, impacts on adult summary index declined from Deming's published results (2009) as his study's cohort grew older by a decade (second bar from the top in Figure 4). Yet, by 2014, Head Start attendees went to school 0.3 years longer than their siblings not attending any preschool. This positive, potentially important impact, however, translated into neither higher college graduation rate nor to significantly higher adulthood earnings.[20]

[Insert Figure 4 here]

C. *Head Start Impacts on School Age Outcomes*

Could Head Start impacts on school age outcomes explain the cohort differences in adult outcomes shown above? For example, are Head Start impacts on achievement generally positive for Deming's cohorts but negative for the complement cohorts? Although a full econometric mediation analysis (e.g., Heckman and Pinto 2015) was not the focus of this article, school age outcomes might nonetheless be considered as potential mediators (e.g., as cognitive or noncognitive inputs) impacting adulthood outcomes. Estimating Head Start impacts on these earlier outcomes could thus be informative about the processes underlying the pattern of later impacts.[21]

[20] Head Start impacts were positive and statistically significant for the subgroup of siblings whose maternal IQ background was 1 sd below the mean: adulthood index (0.38 sd; se = 0.12); educational attainment (0.45 years of schooling completed; se = 0.23); and earnings (0.44 log-points; se = 0.20). See online appendix Table S11.

[21] We conducted this section's analysis as in Deming (2009) and considered identical outcomes and age groups. The full set of estimates were compiled in the online appendix Tables S8-S9 and S12-S13.

As shown in Figure 5, estimates from Deming (2009) and our replication (Deming's cohort) were aligned. For the complement cohort, patterns of impacts on school age outcomes mirrored impacts on the adulthood outcomes: they went in the opposite direction. This was also the case for the Nontests index (-0.149 sd; se = 0.076), with Head Start's impact on the learning disability diagnosis indicator (reverse scaled; -0.042 sd) being statistically significant at the 5 percent level. This is in line with the negative impacts recorded on the Behavioral problem index (reverse scaled; -0.069 sd; se = 0.048). Yet, we could not detect any Head Start impact on 'Grade retention' (for Deming's cohort, the impact was a 7 percentage points lesser chance to be grade retained, significant at the 10 percent level).[22] Regarding the cognitive tests index, the relatively sustained gains generated by Head Start for Deming's cohort (0.106 sd; se= 0.056) did not reflect those for the complement cohort (-0.020; se = 0.063), while equality between the two estimates could not be rejected ($p = 0.243$).[23]

[Insert Figure 5 here]

Deming (2009) recorded Head Start impact fadeout on the cognitive tests index by age period 11-14: from an estimate of 0.145 sd (se = 0.085) by age period 5-6 to one of 0.055 sd (se = 0.062). Instead, for the complement cohort, a fadeout from a small but positive estimate (0.062 sd; se = 0.069) might have occurred faster by age period 7-10 (-0.030 sd; se = 0.060): the difference in impact with Deming's cohort for this age group (0.133 sd, se = 0.060) was of marginal significance ($p = 0.115$). Complement cohort estimates ended at -0.048 sd (se = 0.066) by age period

[22] Other 'Preschool' impact estimates had similar trends on all these 'noncognitive' outcomes (see online appendix Tables S8-S9).

[23] The cognitive tests and behavioral problems indices considered in Figure 5 were scored as the overall average of all corresponding index scores measured from age 5 to 14. We also considered age periods 5-6; 7-10; and 11-14 (see online appendix, Table S12-13). Head Start impact on BPI index were stable across age groups and cohorts and were of similar magnitude as the 5-14 average (Figure 5).

11-14. Finally, the combined cohorts sample faced a similar trend as the complement cohort; overall, impacts approached zero earlier for later Head Start cohorts.

D. *Robustness Checks*

One possible explanation for cohort differences in the estimated impacts of Head Start is that the more recent cohort was more advantaged, and thus less likely to benefit from Head Start. To check if this was the case, a household human capital factor was constructed (alpha = 0.831) by combining standardized measures of maternal, and both grandparents' education levels, maternal AFQT, the natural logarithm of family permanent income, and the CNLSY Home Observation Measurement of the Environment short-form (HOME). This human capital factor was then interacted with indicators 'Head Start' and 'Preschool' from equation (1): the interaction impact estimate on the adulthood summary index, for combined cohorts sample, was not statistically significant (for any value within the range of the household human capital factor).

Another robustness check was conducted by pooling observations from the combined cohorts sample with a different time-range scheme: a first cohort of siblings born before 1983; a second cohort of siblings born between 1982 and 1988; and a third cohort of siblings born after 1987. Systematically, Head Start impacted more favorably the older birth cohorts sample than the two younger ones. We observed a change of sign in the direction of the impact on many of the outcomes considered here: (non)cognitive school age measures; adulthood summary index; educational attainment and earnings.[24]

[24] We also checked for impacts on an alternative earnings variable, taking this time the natural logarithm of the most recent yearly earnings available in CNLSY 2014 survey-round. Head Start impacts were never statistically significant.

Consistent with Anderson's (2008) study of early childhood interventions life cycle impacts, females appeared to have benefited more than males from Head Start across the board of outcomes considered here (see online appendix, Tables S8-S11). Over the extended outcomes adulthood summary index (Deming's cohort), Head Start impact was estimated at 0.226 sd (se = 0.112) for females versus 0.103 sd (se = 0.101) for males. [25] Further, females possibly carried most of Head Start impact on educational attainment with an estimate at 0.343 sd (se = 0.205) against 0.268 sd (se = 0.206) for males (online appendix, Table S11).

In our analytic model (see Equation 1), other 'Preschool' was also included as a within-family predictor of adult outcomes. Considering later and combined cohorts patterns of school age outcomes by age period: for both Head Start and other 'Preschool', impacts on cognitive outcomes were positive at treatment outset (age 5 to 6); whereas impacts changed sign by the next age period (7 to 10) for Head Start siblings, they changed sign over the next age period (11 to 14) for the other 'Preschool' group. Thus, on this school age cognitive outcome, we observed fadeout for both preschool groups, possibly occurring earlier for Head Start participants (see online appendix, model (5) in Table S12).

Moreover, impacts on the Nontest score index (i.e., grade retention and learning disability diagnosis) were unfavorable and statistically significant, for both preschool statuses (online appendix, Table S8). This trend went in the opposite direction for Deming's cohort. Furthermore, for the complement and combined cohort samples, impacts on the Behavior Problems Index were sometimes significant, mostly similar in magnitude, and unfavorable over all age periods across both preschool groups, with estimates slightly more so for other 'Preschool' attendees during the early adolescence period (age 11-14; online appendix model

[25] Although we could not reject equality between these estimates, we detected a statistically significant Head Start favorable impact difference (see online appendix, Table S10) of about 8 percentage points between genders for the 'Idle' individual outcome (i.e., neither working nor in school) for the combined cohorts sample. Similarly, for the complement cohort, females had a 14 percentage points advantage on the 'Crime' outcome (i.e., whether involved with the justice system).

(5) in Table S13). Overall, we could never statistically reject the equality of estimates between Head Start and other ‘Preschool’ status for any of the considered school age and adulthood outcomes for the later cohorts of siblings (online appendix, Tables S8-S11 respective top panels).

To check whether the difference between Deming’s cohort and complement cohort was due to some of the covariates we presented above, we interacted Head Start with an indicator for cohort membership, along with interactions of Head Start with a series of covariates; first, one at a time, then including all interactions. These covariates comprised: pre-treatment index; family human capital index; mother’s age at child’s birth; child’s age at outcome measurement; indicators for gender, whether white/Hispanic or black; and whether maternal AFQT score was one sd below the mean. Had any of these interactions substantially reduced the estimate from Head Start x Cohorts interaction, then these covariates would have explained some of the cross-cohort change in Head Start impacts between the two cohorts. We did not find evidence that this was the case (online appendix, Table S14).

V. Discussion and Conclusion

In this study we replicated and extended Deming’s (2009) evaluation of Head Start impacts over life-cycle skill formation. We found mixed results for Deming’s cohort of siblings, after having extended adulthood individual outcomes with an additional decade of CNLSY data. Second, replicating Deming’s analytic framework on children born to CNLSY mothers after the children in Deming’s cohort revealed contrasting patterns of impacts. In general, for this new cohort, impacts were negative. In fact, for our study of more recent cohorts, Head Start participation might have been detrimental, relative to home care, on non-cognitive and behavioral measures or on the adulthood summary index. Hence, third, combining both cohorts, Head Start impacts on all measured outcomes were small and not statistically significant.

Deming (2009) projected a 0.105 log points Head Start impact on wages. For Deming's cohorts, we estimated from actual data a non-statistically significant smaller impact at 0.067 log points (s.e. = 0.122; estimate measured up to 2014, siblings being between age 28 and 44).[26] Johnson and Jackson (2017), restricting on earlier cohorts of siblings born up to 1976 and using a dynamic complementarity design (i.e., capitalizing on two exogenous sources of variation separated in time), found a larger, more precise estimate: attending Head Start at age 4—i.e., facing an average Head Start spending versus no spending, *coupled* with an average public K-12 spending—boosted earnings of poor children (measured from age 20 to age 50) by 0.099 log points (s.e. = 0.019).[27] These positive earlier estimates generated from a different identification strategy provide convergent evidence that the Head Start estimates from Deming's cohort were not false positives, also underscoring the importance of replicating and explaining potential differences in the estimates in the current study from more recent cohorts.

By and large, our analysis showed heterogeneous impacts from a large-scale early childhood education program. Although our attempts at explaining the change in estimates of the effects of Head Start across cohorts proved inconclusive—and we cannot entirely dismiss the hypothesis that our obtained estimates for Deming's cohort and complement cohort are false positives (Ioannidis, 2005)—changes in counterfactual conditions in household characteristics (e.g., maternal age, income), a substantial increase in spending on means-tested spending between the 1980s and

[26] As described in the results section, for Deming's cohort, attending Head Start versus no preschool yielded a 0.3 year increase of completed schooling. From this, and based on Card's (1999) review on returns to education of about 5 to 10 percent per year of completed schooling, we might have expected to find evidence of an impact on earnings between 1 and 3.5 percent.

[27] That estimate appeared to be sensitive to subsequent K-12 spending level: coupled with a 10% decrease in K-12 spending, the estimate fell to 0.027 log-point (s.e. = 0.028) and was no longer statistically significant, at the 10 percent level. In contrast, with a 10% increase in K-12, the estimate jumped to 0.171 log-point ($p < .01$).

2010s, and possible changes to program quality (e.g., due to a steep and continuous enrollment increase) might still help explain these variations (Ludwig and Phillips 2008).

Beyond shifts in household characteristics, changes in labor market conditions and the return to specific forms of human capital over time might also explain these variations. Given the increase in job polarization and widening of the wage distribution between low and high earners over the past several decades (Autor, Katz, and Kearney, 2006; Goos and Manning, 2007; Autor and Dorn, 2013) it could be the same set of skills encouraged from Head Start participation and acquired by Head Start attendees more favorably rewarded older birth cohorts. These changes in human capital returns over time may have induced individuals from more recent birth cohorts to invest less in educational attainment relative to earlier older birth cohorts. Our findings that Head Start negatively impacted 'Idleness' (i.e. not employed or enrolled in school) and 'Some College Attended' for more recent birth cohorts but not for earlier cohorts lends support to this claim. Overall, this paper suggests that understanding and eliciting pathways of early skill formation with potential subsequent complementarities could be an important priority for basic human capital research and education policies. The novelty of these findings, combined with the possibility of unobserved changes in the selection process into Head Start during this period, necessitate further research on recent cohorts of Head Start attendees using complementary identification strategies.

REFERENCES

Anderson, K., Foster, J., and David Frisvold, D. (2010). Investing in Health: The Long-Term Impact of Head Start on Smoking. *Economic Inquiry* 48(3), 587-602.

Anderson, Michael L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American statistical Association* 103(484): 1481-1495.

Autor, David H., and David Dorn. 2013. "The growth of low-skill service jobs and the polarization of the US labor market." *American Economic Review* 103(5): 1553-97.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2006. "The polarization of the US labor market." *American economic review* 96(2): 189-194.

Bailey, Martha J., Shuqiao Sun, and Brenden Timpe. 2018. "Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency." *Working Paper*.

Barr, Andrew, and Chloe R. Gibbs. 2018. "Breaking the Cycle? Intergenerational Effects of an Anti-Poverty Program in Early Childhood." *Working Paper*.

Bauer, L., and Diane Schanzenbach. 2016. *The Long-term Impact of the Head Start Program*, Policy File.

Card, David. 1999. "The causal effect of education on earnings." In *Handbook of labor economics*, vol. 3, pp. 1801-1863. Elsevier.

Carneiro, Pedro, and Rita Ginja. 2014 "Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start." *American Economic Journal: Economic Policy* 6(4): 135-73.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of*

Economics 126(4): 1593-1660.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. "Interpreting the evidence on life cycle skill formation." *Handbook of the Economics of Education* 1: 697-812

Currie, Janet, and Douglas Almond. 2011. "Human capital development before age five." In *Handbook of labor economics*, Volume 4, 1315-1486. New York: Elsevier.

Currie, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *The American Economic Review* 85(3): 341.

Currie, Janet, and Duncan Thomas. 1999. "Does Head Start help hispanic children?" *Journal of Public Economics* 74(2): 235-262.

Deming, David. 2009. "Early childhood intervention and life-cycle skill development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1 (3): 111-34.

Duncan, Greg J., and Katherine Magnuson. 2013. "Investing in preschool programs." *Journal of Economic Perspectives* 27 (2): 109-32.

Frisvold, David E. 2006. “Head Start Participation and Childhood Obesity.” *SSRN Electronic Journal*.

Frisvold, David E., and Julie C. Lumeng. 2011. “Expanding Exposure: Can Increasing the Daily Duration of Head Start Reduce Childhood Obesity?” *Journal of Human Resources* 46(2), 373-402.

Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-term effects of Head Start." *American economic review* 92(4): 999-1012.

Goos, Maarten, and Alan Manning. 2007. "Lousy and lovely jobs: The rising polarization of work in Britain." *The review of economics and statistics* 89(1): 118-133.

Heckman, James J., and Stefano Mosso. 2014. "The economics of human development and social mobility." *Annual Review of Economics* 6(1): 689-733.

Heckman, James J., and Rodrigo Pinto. 2015. "Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs." *Econometric reviews* 34(1-2): 6-31.

Hoynes, Hilary, Diane Whitmore Schanzenbach, and Douglas Almond. 2016. "Long-run impacts of childhood access to the safety net." *American Economic Review* 106(4): 903-34.

Ioannidis, John PA. 2005. "Why most published research findings are false." *PLoS medicine* 2(8): e124.

Johnson, Rucker C., and C. Kirabo Jackson. 2017. “Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending”. National Bureau of Economic Research (NBER) Working Paper 23489.

Ludwig, Jens, and Douglas L. Miller. 2007. "Does Head Start improve children's life chances? Evidence from a regression discontinuity design." *The Quarterly journal of economics* 122(1): 159-208.

Ludwig, Jens, and Deborah A. Phillips. 2008. "Long-term effects of Head Start on low-income children." *Annals of the New York Academy of Sciences* 1136(1): 257-268.

Miller, Douglas L., Shenhav, Na’ama, and Michel Z. Grosz. 2018. “Selection into Identification in Fixed Effects Models with Application to Head Start.” *Working Paper.*

Thompson, Owen. 2018. “Head Start’s Long-Run Impact: Evidence from the Program’s Introduction.” *Journal of Human Resources* 53(4): 1100-1139.

U.S. Department of Health and Human Services, Administration for Children and Families. 2018. *Head Start Program Information Reports.* Washington, DC. https://hses.ohs.acf.hhs.gov/pir/ (accessed August 14, 2018).

TABLE 1. HOUSEHOLD CHARACTERISTICS AVERAGED OVER HEAD START AND NO PRESCHOOL STATUS

	Head Start			No preschool			Difference HS-None
	Deming's cohort	Complement cohort	Combined cohorts	Deming's cohort	Complement cohort	Combined cohorts	
Permanent income							
Full sample	32,884 [21,810]	39,800 [27,539]	35,970 [24,830]	42,764 [30,000]	61,857 [49,704]	52,445 [41,989]	-0.33/-0.44/ -0.39
Fixed effects subsample	34,672 [25,443]	40,465 [29,118]	37,571 [26,961]	41,587 [27,968]	61,793 [53,019]	53,938 [45,831]	-0.25/-0.40/ -0.36
Mother < high school							
Full sample	0.24 [0.43]	0.12 [0.33]	0.19 [0.39]	0.24 [0.43]	0.12 [0.32]	0.18 [0.38]	0.00/0.00/ 0.03
Fixed effects subsample	0.28 [0.45]	0.14 [0.35]	0.20 [0.40]	0.22 [0.42]	0.12 [0.33]	0.16 [0.37]	0.14/0.06/ 0.11
Mother some college							
Full sample	0.28 [0.45]	0.40 [0.49]	0.33 [0.47]	0.25 [0.43]	0.44 [0.50]	0.34 [0.48]	0.07/-0.08/ -0.02
Fixed effects subsample	0.25 [0.43]	0.42 [0.49]	0.34 [0.47]	0.27 [0.44]	0.43 [0.50]	0.37 [0.48]	-0.05/-0.02/ -0.06
Maternal AFQT							
Full sample	-0.61 [0.61]	-0.50 [0.71]	-0.56 [0.66]	-0.36 [0.80]	0.11 [1.03]	-0.12 [0.95]	-0.31/-0.59/ -0.46
Fixed effects subsample	-0.62 [0.61]	-0.48 [0.70]	-0.54 [0.66]	-0.21 [0.81]	-0.01 [0.99]	-0.18 [0.92]	-0.50/-0.47/ -0.39
Grandmother's education							
Full sample	9.16 [3.09]	9.69 [3.08]	9.39 [3.09]	9.45 [3.23]	10.41 [3.39]	9.92 [3.35]	-0.09/-0.21/ -0.16
Fixed effects subsample	9.13 [3.12]	9.79 [3.00]	9.50 [3.04]	9.69 [3.14]	10.29 [3.30]	10.07 [3.22]	-0.18/-0.15/ -0.18
Sample size	779	637	1,491	1,931	1,857	3,658	
Sample size FE	435	475	972	769	1,098	1,799	

Notes: Means and standard deviations were obtained for the full and sibling fixed effects samples, across cohorts. The difference in means between Head Start and No preschool status (Difference HS-None) were reported for Deming's cohort/Complement cohort/Combined cohort (in that order; in standard deviations units). Permanent income is the average over reported years of household reported net income (in 2014 dollars). The AFQT was age normed based on the NLSY79 empirical age distribution of scores, then standardized (mean = 0; s.d. = 1). Mean differences not reported here, between Deming's cohort and Complement cohort, were all significant at the 1 percent level (see online appendix Table S2).

TABLE 2. HEAD START IMPACTS ON COHORTS' ADULTHOOD SUMMARY INDEX

		Head Start	Other preschool	*p*-value (HS = Other)	R^2
Deming (2009)[a]	(1)	**0.140**	0.083	0.468	0.121
		(0.068)	(0.076)		
Measurement period[b]: 1994-2004	(2)	**0.272**	0.114	0.116	0.590
		(0.080)	(0.084)		
Sample size = 1,251	(3)	**0.228**	0.069	0.080	0.617
[364/364]		(0.072)	(0.072)		
Deming's cohort[c]	(1)	**0.142**	0.079	0.403	0.213
		(0.063)	(0.063)		
Measurement period: 1994-2014	(2)	**0.179**	0.017	0.066	0.639
		(0.072)	(0.092)		
Sample size = 1,251	(3)	**0.166**	0.034	0.134	0.686
[364/364]		(0.069)	(0.066)		
Complement cohort[d]	(1)	**-0.117**	0.025	0.013	0.235
		(0.057)	(0.057)		
Measurement period: 2004-2014	(2)	-0.164	-0.053	0.298	0.711
		(0.102)	(0.079)		
Sample size = 2,144	(3)	**-0.145**	-0.04	0.148	0.728
[497/795]		(0.068)	(0.053)		
p-value for model (3)[e] (Deming's = complement)		0.005	0.644		
Combined cohorts[f]	(1)	-0.024	0.059	0.048	0.239
		(0.04)	(0.034)		
Measurement period: 1994-2014	(2)	-0.008	0.005	0.779	0.612
		(0.005)	(0.035)		
Sample size = 3,738	(3)	-0.011	-0.003	0.859	0.626
[951/1,275]		(0.041)	(0.035)		

Notes: Adulthood summary index (standardized) is a composite of 6 indicators: high school graduation; college attendance; teen-age parenthood; either working or attending school; involvement with the justice system; and poor health status. Model (1): adulthood index is regressed on Head Start and other preschool participation indicators, along with pre-treatment covariates and standardized permanent income; maternal AFQT score; one indicator for maternal high school graduation and one for some college attendance; siblings' gender and age. Model (2): same as model (1) but with family fixed-effect only, no pre-treatment covariates. Model (3): same as model (2) with pre-treatment covariates included. Standard errors are in parenthesis and clustered at the family level. Estimates in bold case were significant at the 5 percent level or less. [a] Deming published results. [b] Outcomes measurement period. [c] For Deming's cohort (compared with Deming 2009), individual outcomes composing the adulthood index were extended up to 2014. [d] Complement cohort includes siblings fitting the same criteria as in Deming (2009) but found eligible from 1990 to 2000. [e] *p*-value = estimates' difference testing between Deming's and complement cohorts' impacts estimated in model (3). [f] Combined cohorts integrates both Deming's and the complement cohorts.

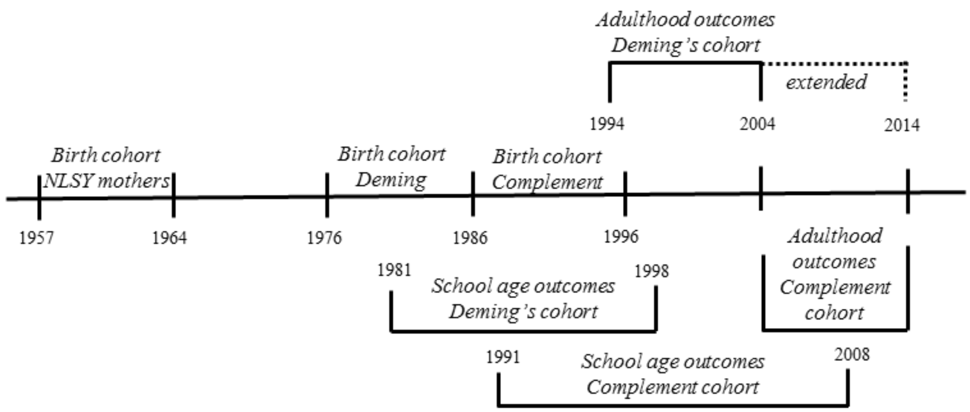

FIGURE 1. BIRTH AND OUTCOMES TIME RANGE BY COHORTS

Notes: Time-wise, the combined cohorts sample (not shown) encompasses Deming's and complement cohorts. Boundary end points are approximate, they include nonetheless the bulk of each distribution: around 95 and 85 percent for Deming's cohort and complement cohort, respectively.

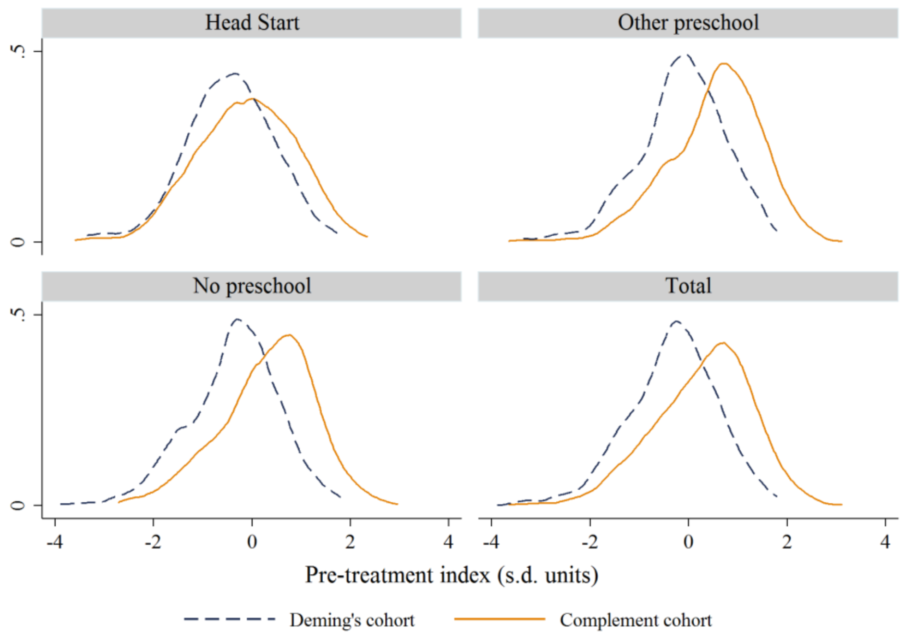

FIGURE 2. PRE-TREATMENT INDEX KERNEL DENSITY ESTIMATION BY PRESCHOOL STATUS ACROSS COHORTS

Notes: Distributions were smoothed, and densities estimated via the Epanechnikov kernel function. Kolmogorov-Smirnov tests all indicated non-equality, at the 0.1 percent level, between compared densities.

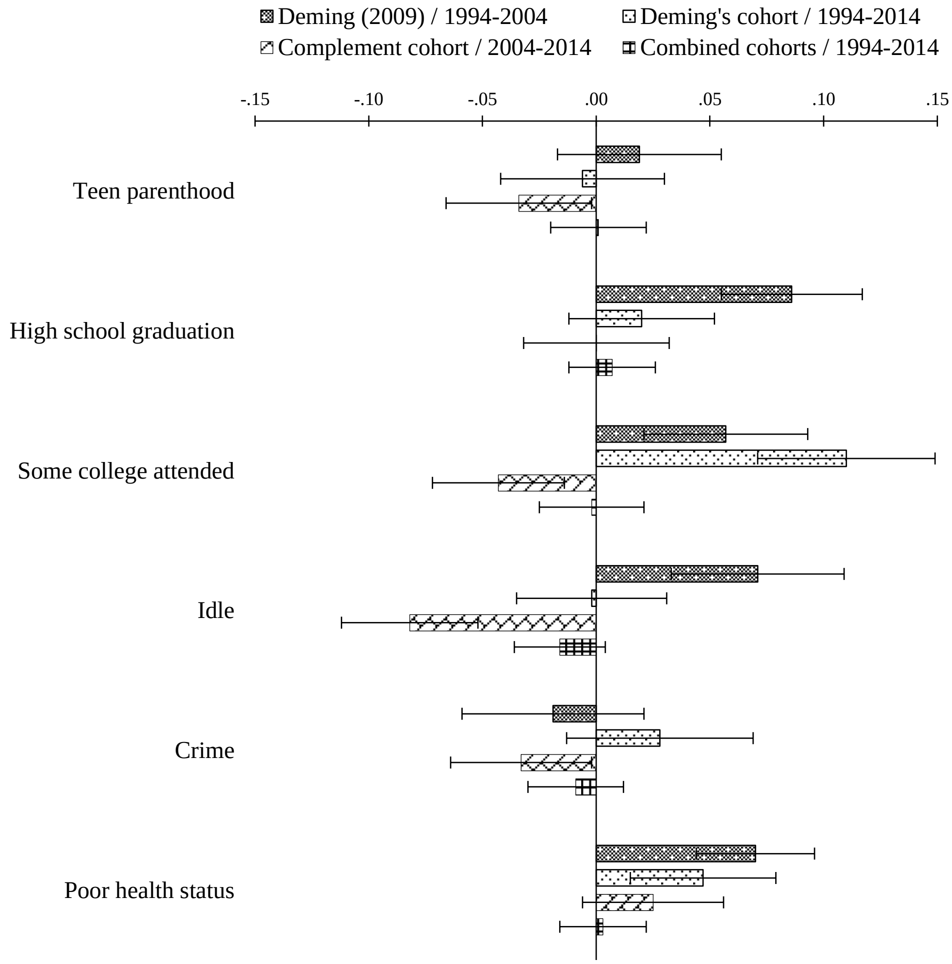

FIGURE 3. HEAD START IMPACTS ON THE ADULTHOOD INDEX INDIVIDUAL OUTCOMES ACROSS COHORTS

Notes: Measurement period are displayed to right of each label. Impacts are expressed as proportions. Deming's cohort, N = 1,251; complement cohort, N = 2,144; combined cohorts, N = 3,768. The counterfactual was a no preschool attendance. Error bars represent standard errors which were clustered at the family level. Estimates were oriented such that a positive value represents a more favorable outcome.

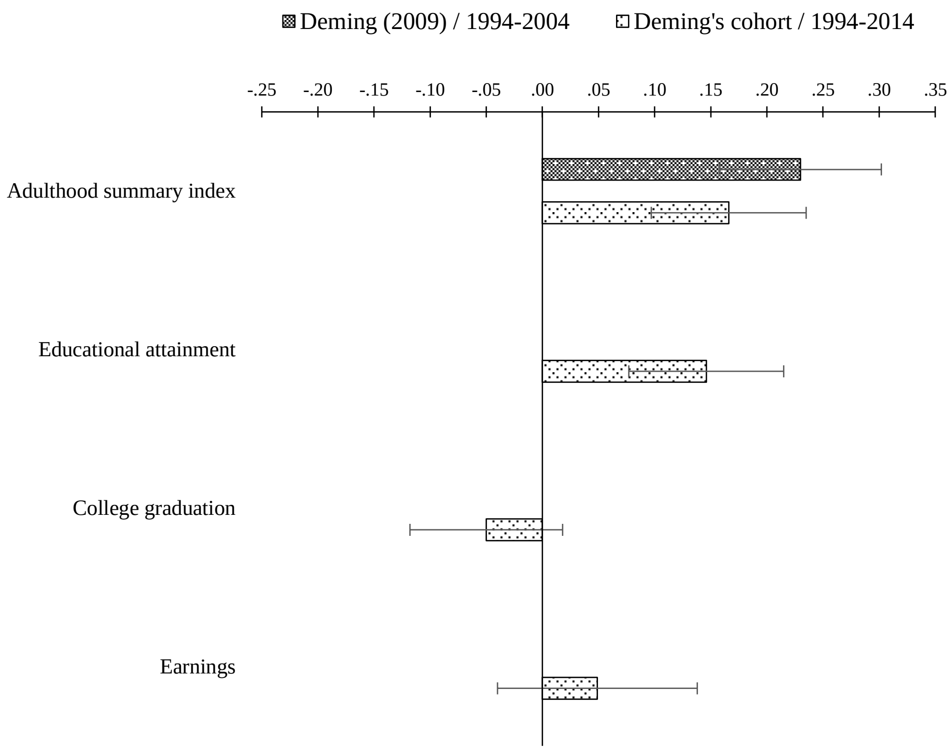

FIGURE 4. HEAD START LONGER-RUN IMPACTS

Notes: Measurement period are displayed to the right of each label. Impacts are expressed in standard deviation units. For both cohorts, N = 1,251. Recall that Deming (2009) study did not estimated Head Start impacts on educational attainment, college graduation and earnings; hence no Deming (2009) bar-estimates for these outcomes. The counterfactual was a no preschool attendance. Error bars represent standard errors which were clustered at the family level.

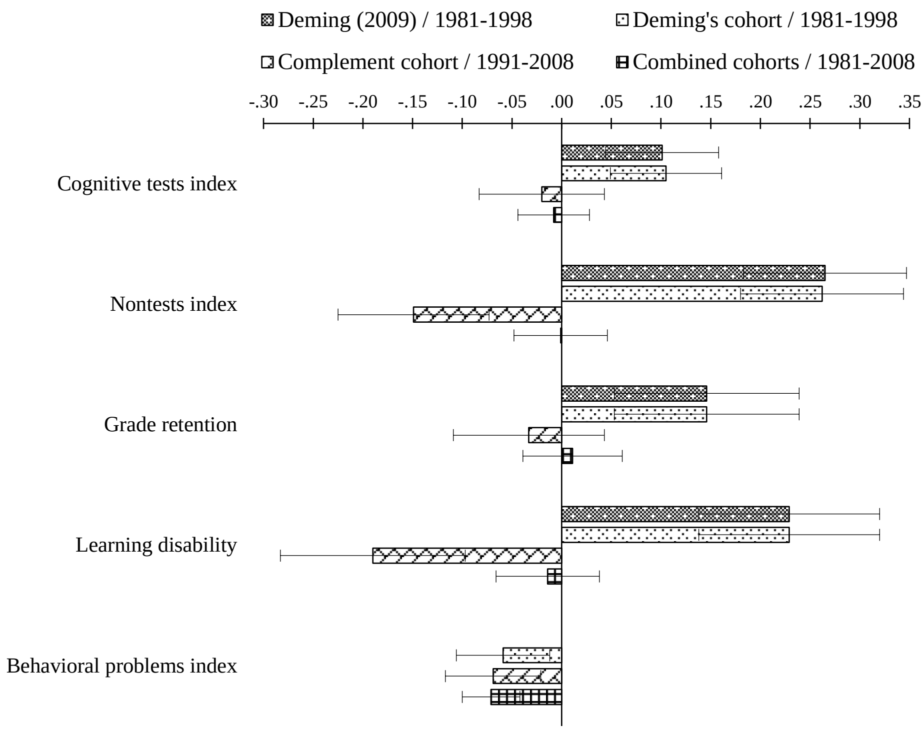

FIGURE 5. HEAD START IMPACTS ON SCHOOL AGE OUTCOMES ACROSS COHORTS

Notes: Measurement period are displayed to the right of each label. Impacts are expressed in standard deviation units. Deming's cohort, N = 1,251; complement cohort, N = 2,144; combined cohorts, N = 3,768. The counterfactual was a no preschool attendance. Error bars represent standard errors which were clustered at the family level. Estimates were oriented such that a positive value represents a more favorable outcome. Deming (2009) estimated impacts on Behavioral problems index (not statistically significant) but did not report them. 'Grade retention' and 'Learning disability' composed the 'Nontests index'.

APPENDIX

The following acronyms are used in this appendix:

ACS = American Community Survey

AMTE = Average Marginal Treatment Effect

ATET = Average Treatment-Effect-On-The-Treated

BPI = Behavioral Problems Index

CAP = Community Action Program

CDS = Child Development Supplement

CESD = Center for Epidemiological Studies Depression

CNLSY = Children of the NLSY 1979 cohort

DID = Difference-in-Difference

FE = Fixed Effects

FFE = Family Fixed Effects

FOS = Federal Outlays System

GTC = Garces et al. (2002)

HS = Head Start

ICPSR = Inter-university Consortium for Political and Social Research

ITT = Intent-to-Treat

IV = Instrumental Variables

NARA = National Archives and Records Administration

NELS = National Education Longitudinal Study

NLSY = National Longitudinal Survey of Youth, 1979 cohort

OEO = Office of Economic Opportunity

PIAT = Peabody Individual Achievement

PIAT-M = Peabody Individual Achievement Math

PIAT-RR = Peabody Individual Achievement Reading Recognition

PIAT-RC = Peabody Individual Achievement Reading Comprehension

PPVT = Picture Peabody Vocabulary Test

RD = Regression Discontinuity

SAIPE = Small Area Income and Poverty Estimates

SEO = Survey of Economic Opportunity

SFR = School Finance Reforms

SEER = Surveillance, Epidemiology, and End Results

SSA = Social Security Administration

Table A1. Selected Quasi-Experimental Studies of Head Start - Adolescence and Later Life Outcomes

Study and Data	Study Design	Results
FAMILY FIXED EFFECTS		
Does Head Start Make a Difference? (Currie and Thomas (1995)) Birth Cohorts: 1970-1986 Data includes NLSY mothers (born 1957-1965) and CNLSY children born to them from 1986, 1988 and 1990 sampling waves. FFE Sample (PPVT): N = 3,477, FFE Sample (Repeated Grade): N = 728, FFE Sample	FFE regression of outcome on controls for child age, gender, first born status, household income at the time child was age 3, mother FE, and indicators for HS participation and other pre-k participation. Focus on childhood outcomes which include PPVT score, grade repetition, measles immunization and height standardized by age and gender.	**African American:** HS (compared to no pre-k), no significant impacts on grade repetition, PPVT scores or height standardized by age and gender, +9.4% (se = 0.03) more likely to get measles immunization **White:** HS (compared to no pre-k) caused +47.3% (se = .12) less likely repeat a grade, +5.9% (se = 1.52) PPVT scores, +8.2% (se = 0.03) more likely to get measles immunization. No significant impact on height standardized by age and gender

(Measles): N = 4,165, FFE Sample (Height by Age/Gender): N = 4,092		**Heterogeneity:** (1) African Americans: +6.8% (se = 1.93) PPVT at base age, -1.3% (se = 0.31) PPVT X Age interaction, no significant Repeat Grade X AFQT interaction. (2) Whites: +6.9 percent (se = 2.40) PPVT at base age, no significant PPVT X Age interaction, +8.3% (se = 0.32) less likely to repeat a grade for a +10% in normalized maternal AFQT
Does Head Start Help Hispanic Children? (Currie and Thomas (1999)) Birth Cohorts: 1970-1988 Data includes NLSY mothers (born 1957-1965) and CNLSY children born to them from 1986, 1988, 1990 and 1992 sampling waves. Sample restricted to Hispanic children aged 5 and older at time of survey. Sample: N = 750	FFE regression of outcome on controls for child's age, gender, first born status, presence of spouse/partner at age 3, mother employed at age 3 and ln(mean household income) (while child age 3-5), family FE, and indicators for HS participation and other pre-k. Outcomes include PPVT, PIAT-M and PIAT-RR test scores in addition to grade repetition.	**Overall:** HS (compared to no pre-k) participation caused +0.22 (t = 2.09) probability of not repeating a grade, + 9.86 (t = 4.06) PPVT percentile score, +5.15 (t = 1.86) PIAT-M percentile score, +3.05 (t = 0.99) PIAT-RR percentile score. No significant impact of other pre-k (compared to no pre-k). Between Hispanic and non-Hispanic white children, HS closes 2/3 of the gap in probability of grade repetition and at least 1/3 of the gap in test scores
Longer-Term Effects of Head Start (Garces, Thomas and Currie (2002)) Birth Cohorts: 1966-1977 Individuals born 1966-1977 birth cohorts in PSID. Focus on adults age 18-30 by 1995. Oldest PSID birth cohorts excluded from sample. FE Sample (Panels A, B and D): N = 1,742, FE Sample (Panel C): N = 728	FFE regression of outcome on controls for year of birth, race and gender, mother FE, and indicators for HS participation and other pre-k participation. Focus on adulthood outcomes which include completion of high school, attendance of some college, ln(earnings) if household member worked, and crime (ever reported booked or charged of a crime)	**Overall:** HS (compared to no pre-k) had no significant impact on high school completion, +9.2% (se = 0.05) on attended some college, and no significant impact on ln(earnings) at age 23-25 and booked/charged with a crime **Subgroup:** (1) African-Americans: no significant impact of HS on high school completion, attended some college, ln(earnings) age 23-25, -0.12 (se = 0.05) less likely to be booked or charged with a crime. (2) Whites: +20.3% (se = 0.10) more likely to complete high school, +28.1% (se = 0.11) more likely to attend some college, no significant impact of HS on ln(earnings) age 23-25 and booked or charged with a crime

Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start (Deming (2009)) Birth Cohorts: 1970-1986 Data includes NLSY mothers born 1957-1965 and CNLSY children born 1970-1986. Most children enrolled in HS from 1984-1990 and over 4 years old by 1990. FE Sample: N = 1,251	FFE regression of outcome on controls gender, age and first-born status, pre-treatment covariates, mother FE, and indicators for HS participation and other pre-k participation. Outcomes include 1) short-term: PPVT, PIAT Math and Reading Recognition tests, grade retention and learning disability and 2) long-term: high school graduation, college attendance, idleness, crime, teen parenthood and self-reported health status, and a summary index of these young adulthood outcomes	**Individual Outcomes:** HS participation (compared to no pre-k) impacted high school graduation +0.09 (se = 0.03), no significant impact of college attendance, -0.07 (se = 0.04) grade repetition, -0.06 (se = 0.02) learning disability, -0.07 (se = 0.04) idleness, no impact on crime or teen parenthood, -0.07 (se = 0.03) poor health **Test Score Index:** HS participation (compared to no pre-k) impacted test scores by +0.15 sd (se = 0.08) (ages 5-6), +0.13 sd (se = 0.06) (ages 7-10) and +0.06 sd (se = 0.06) (ages 11-14). **Non-Test Index:** HS participation (compared to no pre-k) impacted non-cognitive +0.27 sd (se = 0.08), +0.35 sd (se = 0.12) (black), no significant impact for white/Hispanic), +0.39 sd (se = 0.12) (male), no significant impact for female **Adulthood Index:** HS participation (compared to no pre-k) impacted adulthood summary index by +0.23 sd (se = 0.07), +0.24 (se = 0.10) (black), +0.22 sd (se = .10) (white/Hispanic), +0.18 sd (se = 0.10) (male), +0.27 sd (se = 0.11) (female) **Fade-Out:** Large test score fadeout for blacks, none for white/Hispanic
Investing in Health: The Long-Term Impact of Head Start on Smoking (Anderson, Foster and Frisvold (2010)) Birth Cohorts: 1963-1978 Sample (1999): N = 922, Sample (2003): N = 1,005 Data from PSID. Smoking data from 1999 and 2003 surveys for participants aged 21-36.	FFE regression of outcome on controls age, gender, race, birth order, first born status and a battery of family-level pre-treatment covariates, mother FE, and indicators for HS participation and other pre-k participation. Primary outcome of interest is participation in smoking as a young adult (aged 21-36).	**1999 Sample**: HS (compared to no pre-k) participants -17.3% (se = 0.08) less likely to smoke as a young adult. No significant impact of other pre-k (compared to no pre-k). -24.8% (se = 0.10) difference in probabilities of smoking between HS and other pre-k. **2003 Sample:** No significant impact smoking as a young adult for HS (compared to no pre-k) and other pre-k (compared to no pre-k). -19.4% (se = 0.10) difference in probabilities in smoking between HS and

		other pre-k. Difference no longer statistically significant after controlling for educational attainment. **Cost Benefit:** $9,967 PV of smoking reduction (3% discount rate). Avg. costs of HS participant in 2003 was $7,092. Value of smoking reduction 36-141% of HS program costs.
The Long-Term Impact of the Head Start Program Bauer and Schanzenbach (2016) Birth Cohorts: 1970-1990 Data includes NLSY mothers (born 1957-1965) and CNLSY children born to them. Analyzes HS treatment cohorts from 1974-1994. Two samples are used: 1) FE Long-Term: N = 1,439 individuals in 666 families and 2) FE Second Generation: N = 617 individuals in 300 families. Analysis sample limited to respondents 28 years or older in their most recent sample year and did not attrite after 2010 survey year.	FFE regression of outcome on controls for birth year fixed effects, gender, and pre-treatment characteristics. Age at response is also controlled for in self-control and self-esteem regressions. Although not explicitly provided in model details, it is assumed authors also included family fixed effect, indicator for HS and indicator for pre-k in their model specification given their indicated FFE strategy for uncovering HS impacts. Outcomes include high school graduation rates, some college, post-secondary education (i.e. license or certificate, associate's degree, or bachelor's degree), self-control index, self-esteem, and second-generation parenting practices index.	**Overall:** HS (compared to no pre-k) caused +5% ($p<0.1$) high school graduation rates, +12% ($p<0.1$) some higher education, +10% ($p<0.1$) post-secondary completion, +0.15 sd ($p<0.1$) self-control index, +0.15 sd ($p<0.1$) self-esteem index, +0.25 sd ($p<0.1$) positive parenting index. **Subgroup:** For children of mothers with less than high school education: +10% ($p<0.1$) high school graduation, +12% ($p<0.1$) some higher education, +0.3 sd ($p<0.1$) in self-control index, +0.25 sd ($p<0.1$) in self-esteem index. For African American children: +12% ($p<.1$) some higher education, +8% post-secondary completion, +0.35 sd ($p<0.1$) self-control index, +0.20 sd ($p<0.1$) self-esteem index, +0.3 sd ($p<0.1$) positive parenting index. For Hispanic children: +10% ($p<0.1$) high school graduation, +12% ($p<0.1$) some higher education, +14% post-secondary completion
Selection into Identification in Fixed Effects Models, with Application to Head Start. (Miller, Shenhav and Grosz (2018)) [Working Paper] PSID Birth Cohorts: 1966-1987 CNLSY Birth Cohorts: 1970-1986 **PSID**: Individuals born 1966-1977 from PSID (replication of GTC),	FFE regression of adult outcome for a child on controls for individual and family level characteristics, mother FE, and dummy indicators for HS participation and pre-k participation. Weights are used to make sample representative of national population. Standard errors are clustered at mother level.	**PSID Replication:** no impact on high school graduation of HS participation, 0.15 (se = 0.05) effect on some college (white males), no impact on crime, no impact on age 23-25 log earning. On whole, authors find replication matches GTC results. **PSID Extended:** HS leads to no significant improvements in high school completion, 0.12 (se = 0.05) increase in the likelihood of attending some college for white children, no significant effect on

individuals born 1966-1987 from PSID with survey participants from SEO (extension of GTC). Full Sample (replication): N = 3,399, FFE Sample (replication): N = 1,742. Full Sample (extended): N = 7,363, FFE Sample (extended): N = 5,361, FFE Sample (white / extended): N = 2,986 **CNLSY:** All children from CNLSY at least 4 years old by 1990. Identical sample as using in Deming (2009) and linked to NLSY mothers.	Adult outcomes include index of economic sufficiency, index of health outcomes, high school completion, criminal activity, college attendance.	reductions in criminal activity. No significant impacts on several summary indices of long-run economic and health outcomes or college completion. FFE is 50% larger than the AMTE for the representative sample (i.e. FFE produce LATE that overweights larger families) **CNLSY Replication:** HS (compared to no pre-k) lead to +8.5 percentage points in high school graduation ($p<0.01$), -7.2 percentage points in idleness ($p<0.1$), -5.9 percentage points in learning disability ($p< 0.01$) and -6.9 percentage points ($p<0.01$) in reporting poor health. **Reweighted CNLSY:** The reweighted estimate of HS (compared to no pre-k) lead to 40% smaller impact on high school graduation ($p<0.1$), 34% smaller impact on idleness, 4% smaller impact on disability and 25% smaller impact on poor health. Only reweighted estimates of high school graduation are statistically different than FFE estimates ($p<0.1$).
Breaking the Cycle? Intergenerational Effects of an Anti-Poverty Program in Early Childhood (Barr and Gibbs (2018)) [APPENDIX] [Working Paper] Child Birth Cohorts: 1970-1992 NLSY respondents from 1957-1965 birth cohorts and CNLSY children of these mothers. Sample restricted to individuals over 20 by 2012. FE Sample: N = 3,580	[APPENDIX] FFE regression of outcome on controls for child birth order, sex and age, and mother's birth year and age, mother FE, and indicator for HS participation. Robust standard errors clustered at mother's 1979 household. Specification does not mention inclusion of pre-treatment controls or indicator for other pre-k participation. Outcomes include high school graduation, some college, crime, teen parenthood and standardized (mean 0, std 1) index of adulthood outcomes.	[APPENDIX] No significant impact on high school graduation, some college, crime or teen parenthood. +0.26 sd (se = 0.15) impact on index of adulthood outcomes. For African-Americans and "South" region, no significant impact on any outcome. For Male, 19.5% (se = 0.10) effect on high school graduation, -22.0% (se = 0.12) effect on crime and 0.54 sd (se = 0.21) impact on index of adulthood outcomes.

REGRESSION DISCONTINUITY		
Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design (Ludwig and Miller (2007)) Birth Cohorts: 1966-1978 Data includes OEO files from NARA of all federal expenditures for 1967-1980, child mortality from 1973-1983 Vital Statistics, county-level schooling from 1960-2000 decennial censuses, 1990 special tabulation from Census Bureau for schooling attainment by age, race and gender, restricted-use geo-coded NELS	Non-parametric RD regression using local linear regressions with triangle kernel weights. RD cutoff based on which counties were eligible for assistance in applying for Head Start in 1965. Treatment group is the 300 poorest counties in 1965. Control group the next subsequent 300 poorest counties. Focus on average county level outcomes including child mortality rates, high school completion and some college	**Overall:** HS participation increased high school completion rates (18-24, 1990) by 3-4% (p-values <0.05), increased some college (all cohorts) by around 3-5% (p-values < 0.05), and reduced childhood mortality (age 5-9) by 33-50% (p-values < 0.05).
Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start (Carneiro and Ginja (2014)) Birth Cohorts: 1977-1996 Data includes NLSY mothers (born 1957-1965) and CNLSY children born to them. Focus on sample of children whose income between 15% - 185% for relevant income cutoff. Analysis Sample: N = 2,833	Fuzzy RD regression using discontinuities in the probability of participation in HS created by eligibility rules Outcomes include behaviors (e.g. drug use, obesity, grade repetition, alcohol use, school damage, smoking, drinking, sex, special education and BPI), cognitive (e.g. PIAT-M, PIAT-RR, PIAT-RC), crime (i.e. ever convicted or sentenced), CESD, high school diploma, birth control, idle, ever in college, ever work and standardized summary index composed of a weighted average of standardized (mean 0 and std 1) outcome variables.	**<u>Reduced Form ITT</u>** **males ages 20-21:** HS has no significant impact on high school diploma, birth control, ever in college or ever worked. HS reduces the probability of crime by .40 (se = 20), reduces idleness by 0.53 (se = 0.25). Overall, no significant impact on summary index. **males ages 16-17:** HS reduces probability of obesity by 0.47 (se = 0.19), reduces CESD by 0.33 sd (se = 0.10). Overall, +0.16 sd (se = 0.09) impact on summary index. **males ages 12-13:** HS reduces probability of obesity by 0.38 (se = 0.17), reduces BPI by 0.27 sd (se = 0.13), reduces probability that health requires use of special equipment by 0.78 (se = 0.29), reduces probability of need for frequent visits by doctors by 0.32 (se = 0.18).

		No impact on test score measures. Overall, +0.21 sd (se = 0.08) impact on summary index. *Note:* Most effects found for children 12-13 years of age are driven by set of kids attending HS in the 1980s. **Structural Equations** **males ages 20-21:** HS participation had no significant impact on summary index or idleness. HS participation lead to a -22% reduction in crime **males ages 16-17:** HS participation had no significant impact on being overweight and summary index. -0.55 sd in CESD **males ages 12-13:** HS participation lead to -29% in probability of being overweight, -29% in probability of needing special health equipment, +129% sd in summary index (due to very fuzzy 1st stage)
	DID + DID VARIANTS	
Expanding Exposure: Can Increasing the Daily Exposure to Head Start Reduce Childhood Obesity? Birth Cohorts: children of program years spanning 2001-2002 through 2005-2006 (Frisvold and Lumeng (2011)) Data from 2001-2002 through 2005-2006 administrative data from a Michigan HS program. Sample: N = 1,833	First model is DID regression of weight status at end of HS year on weight status at the beginning of the year, controls for individual and family characteristics, indicator for full-day HS participation, and year FE. Second model capitalizes on elimination of full-day HS expansion grants in 2003. 16 full-day classes in 2002 to only four in 2003. Implied IV specification where instrument is % of full-day funded slots. Controls include a battery of family characteristics.	**DID:** Full-day HS reduces obesity by 9.2 percentage points (se = 0.026) (males), 6 percentage points (se = 0.037) (black), no impact for females or white children. Overall, reduces obesity by 3.9 percentage points (se = 0.018) (i.e. akin to change of 20 calories per day with no change in physical activity) **IV:** Full-day HS participation led to a decrease in obesity by 17.6 percentage points. F-stat on excluded instrument is 12.23.
Head Start's Long-Run Impact: Evidence from the Program's Introduction (Thompson (2018))	Regression of adult outcome for a child on race, gender, birth order, number of siblings and maternal education controls, county of birth and cohort FE, and 3-year	**Overall:** 0.125 (se = 0.05) effect of HS exposure on educational attainment, 0.02 (se = 0.01) effect on college graduation, \$2,199 (se = 877.15) effect on own income, \$2,918 (se = 1,437.38) effect on household

Birth Cohort: 1957-1964 Data from 1966–1970 CAP Records from NARA, 1968 and 1972 FOS files NARA (cross-validation), 1965, 1966 and 1968 OEO reports (cross-validation), NLSY respondents born 1957–1964, CNLSY respondents born to NLSY mothers. Analysis Sample: N = 2,685, where sample is restricted to respondents who were ages 2-7 at the time of local Head Start implementation and who did not have a parent who attended college.	average of county-level of HS spending when birth cohort age 3-6. Outcomes include own income, household income, unemployment, educational attainment, high school dropout, college graduate, self-rate health, health conditions, health limitations and standardized adulthood summary index composed of a weighted average of standardized (mean 0 and std 1) outcome variables. Study design measures the effect of exposure to HS funding rather than direct impact of Head Start participation.	income, and -0.05 (se = 0.02) effect on health limitation, and 0.081 sd (se = 0.02) on composite index of child's long-term outcomes, **Subgroup Analyses:** 0.06 sd (se = 0.03) effect of HS exposure on composite index of child's long-term outcomes for whites, 0.09 sd (se = 0.04) effect of HS exposure on composite index of child's long-term outcomes for blacks
Breaking the Cycle? Intergenerational Effects of an Anti Poverty Program in Early Childhood (Barr and Gibbs (2018)) [Working Paper] Child Birth Cohorts: 2nd generation 1966–1968 CAP Records and 1968–1980 FOS Files from NARA for HS availability in FY 1966-1968, NLSY respondents born 1960–1964, CNLSY respondents born to NLSY mothers. Full sample: N = 3,533 (2nd gen child) and N = 2,398 (1st gen mother). High Impact sample: N = 1,687 (2nd gen child) and N = 821 (1st gen mother). Low Impact sample: N = 2,732 (2nd gen child) and N = 1,398 (1st gen mother)	Regression of adult outcome for a child on sex, age, age squared, race and mother's birth order controls, county of birth and birth year FE, and indicator variable for HS availability for a mother by birth cohort and county of birth. Outcomes include teen parenthood, interaction with criminal justice system (arrests, convictions or probations), high school graduation, some college attendance, and a summary index composed of a weighted average of standardized (mean 0 and std 1) outcome variables. Heterogeneity tests by race, gender and geographic location. Study design measures the effect of exposure to HS funding rather than direct impact of Head Start participation.	**High Impact Sample:** 0.67 (se = 0.23) years of educational attainment, +12.7% (se = 0.05) on high school graduation, 16.9% (se = 0.06) effect on some college, -8.6% (se = 0.03) effect on teen pregnancy, -15.6% (se = 0.04) effect on crime, 0.47 sd (se = 0.10) effect on index of child's adulthood outcomes. Max educational attainment not included in main results as many individuals not yet finished with their education. **Low Impact Sample:** 0.28 (se = 0.18) years of educational attainment, 6.4% (se = 0.04) effect on high school graduation, 7.0% (se = 0.05) effect on some college, -5.8% (se = 0.03) effect on teen pregnancy, -6.3% (se = 0.03) effect on crime, 0.22 (se = 0.09) effect on index of child's adulthood outcomes. Max educational attainment not included in main results as many individuals not yet finished with their education. **Subgroup Analyses:** Larger effect on crime and smaller effects on teen parenthood for male children

Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency Bailey, Sun & Timpe (2018) [Working Paper] Birth Cohorts: 1950-1980 Links long-form 2000 Census and 2001-2013 ACS with SSA's Numident file. Focus on individuals age 25-54 years old.	Event study framework where adult outcome of a child regressed on county-level controls, FEs for county of birth, year, and state-by-birth year, HS indicator interacted with child's school age at time of HS's launch. To improve precision and test formally for trend breaks, paper uses a three-part spline specification with knots at ages 6 and -1. Outcomes include summary measures of human capital and economic self-sufficiency. Human capital index composed of high school or GED, some college, 4-year college degree, professional or doctoral degree, years of schooling, and indicator for occupation. Self-sufficiency composed of binary indicators of employment, poverty status, income from public sources, family income, and income from other non-governmental sources; continuous measures of weeks worked, usual hours worked, the log of labor income, log of other income from non-governmental sources, and log ratio of family income to the federal poverty threshold.	**Overall:** Finds HS caused 0.29 year (ATET 95% CI [.14, .49]) increase in schooling, 2.1% (ATET 95% CI [0.005, 0.038]) increase in high-school completion, 8.7% [ATET 95% CI [0.027, 0.092]) increase in college enrollment, 19% (ATET 95% CI [0.025, 0.094]) increase in college completion, and 10% of a standard deviation increase in the adult human capital index. Also finds HS caused nearly 4% of a standard deviation increase in economic self-sufficiency index. Largely driven by 12% reduction in adult poverty and 29% reduction in public assistance receipt. No statically significant impacts of HS on incarceration. **Heterogeneity:** Differential impact of HS on human capital index by Medicaid exposure (ITT $F = 4.1$, $p<0.05$), CHC exposure (ITT $F = 9.3$, $p<0.01$), Food Stamps exposure (ITT $F = 4.5$, $p<0.05$), predicted economic growth (ITT $F = 11.9$, $p<0.01$). Differential impact of HS on self-sufficiency index by Food Stamps exposure (ITT $F = 3.5$, $p<0.06$).
Reducing Inequality Through Dynamic Complementarity: Evidence from Head Start and Public School Spending (Johnson and Jackson (2018)) Birth Cohorts: 1950-1976	First model uses DID instrumental variables (DID-2SLS) where within-county, across-cohort DID variation in HS spending is exploited. Both public K12 spending and the interaction between HS spending per poor 4-year old and public K12 spending are	**HS Spending (DID-2SLS):** For poor children, +0.08 (se = 0.02) years of education, +2.5% (se = 0.007) likelihood of high school graduation, 2.3% (se = 0.005) higher wages (ages 20-50), +0.6% (se = 0.003) less likelihood of adult incarceration, +1.8% (se = 0.005) less likelihood of adult poverty. No significant impact for non-poor children.

Data includes annual HS spending at county level from NARA, ICPSR and SEER, annual public K12 spending at school district level, SFR database, 1968-2015 PSID data for individual long-term outcomes, 1960 Census data for county-level characteristics, and a multitude of datasets to capture info on timing of other key policy changes	instrumented using 2SLS. Second model is uses DID instrumental variables where all where all spending variables are instrumented (2SLS-2SLS). This includes HS spending, public K12 spending and their interaction. Outcomes include educational attainment, high school graduation, wages, incarceration and adult poverty.	**Public K12 Spending (DID-2SLS):** Large and significant positive impacts for poor and non-poor children across probability of high school graduation, years of completed education, wages (age 20-50), less likelihood of adult incarceration and less likelihood of adult poverty. **Dynamic Complementarity (DID-2SLS):** For poor children exposed to typical HS center, a 10% increase in K12 spending leads to +0.59 (se = 0.12) years of education, +14.8% (se = 0.02) likelihood of high school graduation, +17.1% (se = 0.04) higher wages (ages 20-50), +4.7% (se = 0.01) less likelihood of adult incarceration, +12.2% (se = 0.04) less likelihood of adult poverty
INSTRUMENTAL VARIABLES		
Head Start Participation and Childhood Obesity (Frisvold (2006)) Birth Cohorts: ages 5-19 by 2002 Data from PSID, PSID CDS and U.S. Census Bureau's SAIPE. Sample: N = 2,301, Sub-Sample (white): N = 1,138, Sub-Sample (black): N = 973	IV regression using a bivariate probit model where Head Start participation is instrumented by the number of spaces available in a community. Outcomes include childhood overweight and obesity (ages 5-19	**Overall:** No significant impact of HS on obesity (ages 5-19). Significant impact of HS on overweight (ages 5-19) [ATT: -0.25, se = 0.08] **Black:** Significant impact of HS on overweight (ages 5-19) [ATT: -0.334, se = 0.157] and obesity (ages 5-19) [ATT: -0.332, se = 0.158] **White:** No significant impact of HS on overweight and obesity (ages 5-19).

ONLINE APPENDIX

S1. REVIEW OF HEAD START IMPACTS ON SHORT AND MEDIUM-TERM OUTCOMES

What are the impacts of Head Start for children at the end or soon after program completion? Morris et al. (2018) found in a recent reanalysis of the Head Start Impact Study (HSIS; a 2002 national randomized evaluation of Head Start), results consistent with two other HSIS reanalyzes, similarly considering counterfactual types (Zhai, Brooks-Gunn, and Waldfogel 2014; Kline and Walters 2016). All three studies concluded that attending Head Start vs. no preschool yielded short benefits on cognitive test scores (measured 1 year after treatment) which quickly faded during the elementary school years. Morris et al., however, also found that centers' quality, geographic locations, counterfactual care alternatives, and dual language learner status, *all* moderated the treatment effect for some cognitive outcomes (e.g., receptive vocabulary; early numeracy). The authors thus emphasized interpreting estimated impacts in light of counterfactual and contextual condition comparisons. Finally, Kline and Walters projected impacts of test scores onto adulthood earnings gains and—after accounting for program substitution fiscal externalities—derived a greater than 1 benefit-cost ratio.

Accounting for previous Head Start evaluations' heterogeneous results, Shager et al.'s meta-analysis (2013) covering 28 Head Start studies, showed that about 41percent of cross-study differences were accounted by the quality of the evaluation research design. Lesser quality being mostly associated with *not* including an indicator distinguishing between ‘active’ (other preschool participation) and ‘passive’ (no preschool) control group. The authors summarized the effect sizes at 0.27 sd for cognitive and achievement outcomes measured less than a year after treatment. They also found that when the control group attended some of form of preschool, effect sizes on the same outcomes were smaller at 0.08 sd and not statistically significant.

Shager and colleagues (2013) did not specify for the kind of instruction delivered to the active control group. Instead, they considered whether outcomes of interest (i.e., skills) responded to any instruction. Camilli et al. (2010) conducted a meta-analysis spanning 128 center-based early intervention evaluations (about a fourth of which on Head Start programs) conducted from 1960 to 2000. They found that an individualized and explicit teacher-led form of instruction yielded an immediate boost on cognitive outcomes compared to a student-directed learning structure. Their overall preschool impact estimation of 0.23 sd on cognitive outcomes (versus no-preschool) was in line with Shager et al.’s results.

Equally consistent with other previous findings on the HSIS (e.g., Puma et al. 2012; Walters 2015), Camilli et al. (2010) found that early impacts tended to fade rapidly after intervention: this was especially the case when contrasting with a passive counterfactual group.[29] However, the authors could not reject the

[29] Fadeout was less pronounced when comparing to a preschool alternative control group, which suggested that fadeout was proportionally related to corresponding impact magnitude. It was much lower in the case of preschool alternatives counterfactual.

persistence into middle school (at or after age 10) of positive impacts on school and socio-emotional multiple outcomes. Deming (2009) found similar patterns of Head Start early cognitive boosts vanishing by age 11, along a persistent favorable impact (0.27 sd) on a non-test score composite of grade retention and learning disability diagnosis measured by age 14. Neither Walters's nor Deming's studies reported significant impacts on psychometric measures of socio-emotional outcomes.

Investigating the relations between Head Start characteristics inputs and medium-term effectiveness (HSIS followed children up to third grade), Walters (2015) tested for interactions between seven center characteristic indicators and Head Start participation. Among these, only interactions with the variable 'delivery of a full day of service' (0.138 sd; se = 0.055), and 'any staff having a teaching license' (0.127 sd = 0.068) were marginally impactful on cognitive outcomes. This last interaction, however, had a potentially moderate impact (up to 0.22 sd, within the 95 percent confidence interval), in a multivariate model including all interactions at once. However, estimates were not robust when applying the analysis to medium-run ones. For Walters, this indicated that notwithstanding program characteristics inputs' moderation of short-run impacts, the latter faded out all the same.

From one perspective, given the well-known impacts in adulthood of small RCTs of early childhood education programs targeted at economically disadvantaged children in the 1960s (see Elango et al., 2016 for review), the short-run benefits observed in the HSIS and Deming's sibling comparison study forecasted beneficial effects in adulthood for children who attended Head Start. In their cost-benefit analysis of Head Start, Kline and Walters (2016; and reference therein) noted that other educational interventions have yielded positive adult outcomes despite fadeout of test score gains. Deming's cost-benefit analysis projected earnings based on his estimates of Head Start effects on positive outcomes in early adulthood, which circumvents the problem of projecting long-term effects on the basis of short-term effects.

From a different perspective, the pattern of full fadeout of effects of Head Start on measures of academic achievement ran in notable contrast to findings from both the Perry Preschool and Abecedarian programs, both of which produced persistent impacts on academic achievement into participants' 20s (Campbell et al. 2002; Schweinhart et al. 2005). Therefore, the underlying developmental processes may not have been causally impacted in the same way in recent Head Start cohorts as in these classic early childhood education studies, and therefore may yield different patterns of adult impacts.

REFERENCES

Camilli, Gregory, Sadako Vargas, Sharon Ryan, and W. Steven Barnett. 2010. "Meta-analysis of the effects of early education interventions on cognitive and social development." *Teachers college record* 112(3): 579-620.

Campbell, Frances A., Craig T. Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson. 2002. "Early childhood education: Young adult outcomes from the Abecedarian Project." *Applied developmental science* 6(1): 42-57.

Elango, Sneha, Jorge Luis García, James J. Heckman, and Andrés Hojman. 2015. “Early childhood education”. National Bureau of Economic Research Working Paper 21766.

Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating public programs with close substitutes: The case of Head Start." *The Quarterly Journal of Economics* 131(4): 1795-1848.

Morris, Pamela A., Maia Connors, Allison Friedman-Krauss, Dana Charles McCoy, Christina Weiland, Avi Feller, Lindsay Page, Howard Bloom, and Hirokazu Yoshikawa. 2018. "New findings on impact variation from the Head Start Impact Study: Informing the scale-up of early childhood programs." *AERA Open* 4(2).

Puma, Mike, Stephen Bell, Ronna Cook, Camilla Heid, Pam Broene, Frank Jenkins, Andrew Mashburn, and Jason Downer. 2012. "Third Grade Follow-Up to the Head Start Impact Study: Final Report. OPRE Report 2012-45." *Administration* *for Children & Families*.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, and Milagros Nores. 2005. "Lifetime effects: The high/scope Perry preschool study through age 40 (monographs of the high/scope educational research foundation, 14)." *Ypsilanti, MI: High Scope Educational Research Foundation*.

Shager, Hilary M., Holly S. Schindler, Katherine A. Magnuson, Greg J. Duncan, Hirokazu Yoshikawa, and Cassandra MD Hart. 2013. "Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes." *Educational Evaluation and Policy Analysis* 35(1): 76-95.

Walters, Christopher R. 2015. "Inputs in the production of early childhood human capital: Evidence from Head Start." *American Economic Journal: Applied Economics* 7(4): 76-102.

Zhai, Fuhua, Jeanne Brooks-Gunn, and Jane Waldfogel. 2014. "Head Start's impact is contingent on alternative type of care in comparison group." *Developmental psychology* 50(12): 2572.

Table S2. Household Selected Characteristics by Preschool Status Across Cohorts

	Head Start				Preschool				None				
	Deming's cohort	Complement cohort[a]	Combined cohorts[b]	*p*-value[c]	Deming's cohort	Complement cohort	Combined cohorts	*p*-value	Deming's cohort	Complement cohort	Combined cohorts	*p*-value	HS–None[d]
Permanent	32,884	39,800	35,970	<.001	62,076	92,664	80,810	<.001	42,764	61,857	52,445	<.001	-.39
income	[21,810]	[27,539]	[24,830]		[41,520]	[107,713]	[89,545]		[30,000]	[49,704]	[41,989]		
Fixed effects	34,672	40,465	37,571	.002	49,165	77,243	72,957	<.001	41,587	61,793	53,938	<.001	-.36
subsample	[25,443]	[29,118]	[26,961]		[29,600]	[62,153]	[55,368]		[27,968]	[53,019]	[45,831]		
Mother < high	.24	.12	.19	<.001	.08	.04	.06	<.001	.24	.12	.18	<.001	.03
school	[.43]	[.33]	[.39]		[.27]	[.20]	[.23]		[.43]	[.32]	[.38]		
Fixed effects	.28	.14	.20	<.001	.12	.06	.08	.001	.22	.12	.16	<.001	.11
subsample	[.45]	[.35]	[.40]		[.32]	[.24]	[.28]		[.42]	[.33]	[.37]		
Mother some	.28	.40	.33	<.001	.45	.64	.57	<.001	.25	.44	.34	<.001	-.02
college	[.45]	[.49]	[.47]		[.50]	[.48]	[.50]		[.43]	[.50]	[.48]		
Fixed effects	.25	.42	.34	<.001	.32	.56	.48	<.001	.27	.43	.37	<.001	-.06
subsample	[.43]	[.49]	[.47]		[.47]	[.50]	[.50]		[.44]	[.50]	[.48]		
Maternal AFQT	-.61	-.50	-.56	.001	.01	.49	.31	<.001	-.36	.11	-.12	<.001	-.46
	[.61]	[.71]	[.66]		[.85]	[1.03]	[1.00]		[.80]	[1.03]	[.95]		
Fixed effects	-.62	-.48	-.54	.001	-.21	.19	.03	<.001	-.21	-.01	-.18	<.001	-.39
subsample	[.61]	[.70]	[.66]		[.81]	[1.02]	[.97]		[.81]	[.99]	[.92]		
Grandmother's	9.16	9.69	9.39	.001	10.69	11.40	11.13	<.001	9.45	10.41	9.92	<.001	-.16
education	[3.09]	[3.08]	[3.09]		[3.09]	[3.06]	[3.00]		[3.23]	[3.39]	[3.35]		
Fixed effects	9.13	9.79	9.50	.001	10.10	10.78	10.51	<.001	9.69	10.29	10.07	<.001	-.18
subsample	[3.12]	[3.00]	[3.04]		[3.06]	[3.22]	[3.13]		[3.14]	[3.30]	[3.22]		
Sample size	779	637	1,491		994	1,681	2,724		1,931	1,857	3,658		
Sample size FE	435	475	972		459	835	1,349		769	1,098	1,799		

Notes. [a] Complement cohort includes children deemed eligible from 1992 to 2000. [b] [c] *p*-value corresponds to estimates difference tests between Deming's cohort and Complement cohort, both included in the Combined cohorts. [d] Combined cohort difference between Head Start vs. None (in sd units). Permanent income is in 2014 dollars. Standard deviations are in brackets.

Table S3. Household Selected Characteristics by Preschool Status Across Cohorts (White/Hispanic)

	Head Start					Preschool					None					
	Deming (2009)[A]	Deming's cohort	Complement cohort[a]	Combined cohorts[b]	p-value[c]	Deming (2009)	Deming's cohort	Complement cohort	Combined cohorts	p-value	Deming (2009)	Deming's cohort	Complement cohort	Combined cohorts	p-value	HS–None[e]
Permanent income[d]	26,553 [19,555]	27,674 [19,022] 34,770 [23,900]	42,645 [30,648]	38,531 [27,449]	<.001	52,130 [34,577]	54,771 [34,410] 68,816 [43,233]	101,193 [116,519]	88,968 [97,500]	<.001	35,592 [23,460]	36,848 [23,877] 46,298 [30,000]	66,956 [52,278]	57,139 [44,583]	<.001	-.42
Fixed effects subsample	27,560 [22,902]	28,479 [21,805] 35,783 [27,397]	45,261 [32,866]	40,150 [30,177]	.001	41,882 [22,403]	43,715 [22,852] 54,926 [28,713]	87,466 [67,507]	72,957 [55,368]	<.001	35,901 [23,600]	37,150 [23,569] 46,673 [29,614]	69,387 [57,066]	61,637 [50,400]	<.001	-.43
Mother < high school	.51 [.50]	.35 [.48]	.16 [37]	.26 [.44]	<.001	.18 [.38]	.08 [.26]	.04 [.20]	.05 [.23]	<.001	.42 [.49]	.27 [.44]	.12 [.33]	.20 [.40]	<.001	.15
Fixed effects subsample	.53 [.50]	.34 [.47]	.18 [.39]	.27 [.44]	<.001	.25 [.43]	.10 [.31]	.06 [.24]	.08 [.27]	.014	.41 [.49]	.25 [.43]	.13 [.34]	.17 [.38]	<.001	.26
Mother some college	.22 [.41]	.22 [.41]	.32 [.47]	.26 [.44]	.004	.41 [.49]	.42 [49]	.65 [.48]	.56 [.50]	<.001	.23 [.42]	.23 [.42]	.44 [.50]	.34 [.47]	<.001	-.17
Fixed effects subsample	.16 [.37]	.16 [.36]	.33 [.47]	.24 [.43]	<.001	.31 [.46]	.32 [.47]	.56 [.50]	.46 [.50]	<.001	.22 [.41]	.23 [.43]	.41 [.49]	.35 [.48]	<.001	-.23
Maternal AFQT	-.44 [.73]	-.45 [.71]	-.31 [.80]	-.39 [.75]	.015	.23 [.85]	.19 [.82]	.68 [1.00]	.50 [.97]	<.001	-.21 [.86]	-.23 [.84]	.27 [1.05]	.03 [.98]	<.001	-.43
Fixed effects subsample	-.48 [.70]	-48 [.69]	-.27 [.77]	-.37 [.73]	.003	.02 [.83]	-.01 [.80]	.42 [1.02]	.25 [.98]	<.001	-.20 [.82]	-.22 [.80]	.18 [1.02]	.02 [.96]	<.001	-.41
Grandmother's education	8.53 [3.50]	8.53 [3.50]	8.73 [3.57]	8.63 [3.52]	.469	10.62 [2.92]	10.62 [2.92]	11.49 [3.18]	11.17 [3.13]	<.001	9.34 [3.36]	9.35 [3.36]	10.36 [3.60]	9.87 [3.52]	<.001	-.35
Fixed effects subsample	8.51 [3.42]	8.51 [3.42]	8.96 [3.54]	8.73 [3.45]	.175	10.09 [3.19]	10.09 [3.19]	10.81 [3.47]	10.52 [3.36]	.002	9.54 [3.34]	9.54 [3.34]	10.25 [3.55]	10.04 [3.44]	<.001	-.38
Sample size	364	364	291	682		745	745	1,359	2,148		1,374	1,380	1,482	2,795		
Sample size FE	229	229	211	478		315	315	619	985		510	510	827	1,269		

Notes. Standard deviations are in brackets. [A] Deming's (2009) published estimates; [a] Complement cohort includes children deemed eligible from 1992 to 2000. [c] *p*-value corresponds to mean difference tests between Deming's cohort and complement cohort. [d] Permanent (child's family) income in the Deming's cohort column is first in 2004 dollars; and below in 2014 dollars. [e] Combined cohorts difference between Head Start status vs. None (in sd units).

Table S4. Household Selected Characteristics by Preschool Status Across Cohorts (Black)

	Head Start					Preschool					None					
	Deming (2009)[A]	Deming's cohort	Complement cohort[a]	Combined cohorts[b]	*p*-value[c]	Deming (2009)	Deming's cohort	Complement cohort	Combined cohorts	*p*-value	Deming (2009)	Deming's cohort	Complement cohort	Combined cohorts	*p*-value[c]	HS–None[e]
Permanent income[d]	24,005 [16,103]	24,856 [15,661] 31,231 [19,678]	37,406 [24,412]	33,810 [22,174]	<.001	32,470 [21,939]	33,353 [18,639] 41,907 [27,398]	56,668 [40,940]	50,387 [36,069]	<.001	25,980 [18,496]	26,993 [18,639] 33,914 [23,419]	41,704 [30,449]	37,243 [27,097]	<.001	-.13
Fixed effects subsample	26,010 [19,559]	26,612 [18,370] 33,436 [23,081]	36,632 [25,147]	35,076 [27,097]	.157	28,940 [22,853]	29,102 [21,983] 36,564 [27,620]	47,947 [26,878]	43,238 [27,269]	<.001	24,164 [16,314]	25,128 [16,815] 31,572 [21,128]	38,615 [27,247]	35,504 [23,829]	.001	-.02
Mother < high school	.33 [.47]	.15 [.36]	.09 [.29]	.13 [.33]	.009	.20 [.40]	.08 [.27]	.05 [.22]	.06 [.23]	.136	.38 [.49]	.17 [.37]	.09 [.28]	.13 [.34]	<.001	.00
Fixed effects subsample	.39 [.49]	.21 [.41]	.11 [.31]	.14 [.35]	.003	.27 [.45]	.14 [.35]	.07 [.26]	.09 [.29]	.045	.37 [.48]	.19 [.39]	.08 [.27]	.14 [.34]	<.001	.00
Mother some college	.31 [.46]	.34 [.48]	.47 [.50]	.40 [.49]	.001	.50 [.50]	.53 [.50]	.60 [.49]	.58 [.49]	.102	.28 [.45]	.30 [.46]	.47 [.50]	.37 [.48]	<.001	.06
Fixed effects subsample	.32 [.47]	.35 [.48]	.49 [.50]	.44 [.50]	.002	.42 [.50]	.48 [.50]	.56 [.50]	.53 [.50]	.111	.30 [.46]	.35 [.48]	.49 [.50]	.41 [.49]	.001	.06
Maternal AFQT	-.75 [.49]	-.75 [.47]	-.66 [.57]	-.71 [.53]	.004	-.51 [.72]	-.52 [.71]	-.33 [.76]	-.41 [.74]	.003	-.68 [.60]	-.68 [.59]	-.54 [.65]	-.62 [.63]	.001	-.14
Fixed effects subsample	-.77 [.48]	-.78 [.47]	-.64 [.59]	-.69 [.55]		-.63 [.66]	-.63 [.65]	-.45 [.66]	-.56 [.63]	.011	-.76 [.56]	-.77 [.54]	-.56 [.65]	-.66 [.59]	<.001	-.05
Grandmother's education	9.71 [2.56]	9.71 [2.56]	10.51 [2.30]	10.02 [2.50]	<.001	10.88 [2.68]	10.88 [2.68]	10.98 [2.44]	10.98 [2.49]	.636	9.70 [2.87]	9.70 [2.87]	10.61 [2.40]	10.09 [2.72]	<.001	-.03
Fixed effects subsample	9.82 [2.59]	9.82 [2.59]	10.44 [2.27]	10.25 [2.34]	.006	10.13 [2.76]	10.13 [2.76]	10.71 [2.36]	10.48 [2.40]	.035	9.98 [2.67]	9.98 [2.67]	10.43 [2.42]	10.14 [2.61]	.042	.04
Sample size	415	415	346	809		249	249	322	576		551	551	375	863		
Sample size FE	206	206	264	494		144	144	216	364		259	259	271	530		

Notes. Standard deviations are in brackets. [A] Deming's (2009) published estimates; [a] Complement cohort includes children deemed eligible from 1992 to 2000. [c] *p*-value corresponds to estimates difference tests between Deming's cohort and complement cohort, both included in the [b] Combined cohorts. [d] Permanent (child's family) income in the Deming's cohort column is first in 2004 dollars; and below in 2014 dollars. [e] Combined cohorts difference between Head Start status vs. None (in sd units).

Table S5. Sibling Differences on Pre-Treatment Covariates by Preschool Status

	Head Start		Preschool		No preschool (Control mean)	
	Complement cohort	Combined cohorts	Complement cohort	Combined cohorts	Complement cohort	Combined cohorts
Pre-treatment index	-0.001	0.005	0.031	0.018	-0.003	-0.031
	(0.063)	(0.035)	(0.039)	(0.027)	[0.975]/2,144	[0.983]/3,738
Attrition	**-0.029**	-0.009	**-0.023**	-0.011	0.045	0.025
	(0.013)	(0.007)	(0.012)	(0.007)	[0.207]/2,215	[0.169]/3,826
Age in 2014 (in years)	**1.07**	0.085	**1.80**	**0.415**	24.45	29.03
	(0.425)	(0.272)	(0.270)	(0.212)	[3.95]/2,144	[5.75]/3,738
PPVT at age 3	1.58	1.78	-0.535	-1.01	22.00	21.05
	(3.38)	(1.99)	(3.73)	(2.05)	[13.25]/396	[12.60]/644
Log of birth weight	-0.008	0.023	**-0.025**	-0.010	4.76	4.73
	(0.018)	(0.012)	(0.012)	(0.009)	[0.203]/1,894	[0.225]/3,449
Very low BW	0.006	-0.007	0.010	0.005	0.006	0.013
	(0.010)	(0.007)	(0.007)	(0.004)	[0.074]/1,894	[0.114]/3,449
In mother's HH, age 0-3	0.003	0.008	-0.002	-0.001	0.988	0.974
	(0.007)	(0.007)	(0.005)	(0.005)	[0.085]/2,123	[0.119]/3,601
Previous health limitation	0.001	0.008	**-0.045**	-0.020	0.095	0.066
	(0.026)	(0.014)	(0.018)	(0.012)	[0.293]/2,123	[0.249]/3,601
Firstborn	0.100	-0.016	**0.277**	0.051	0.183	0.319
	(0.052)	(0.031)	(0.039)	(0.026)	[0.387]/2,144	[0.466]/3,738
Male	-0.019	-0.004	-0.054	-0.010	0.528	0.512
	(0.043)	(0.026)	(0.115)	(0.023)	[0.499]/2,144	[0.500]/3,738
HOME score at age 3	0.829	-0.247	**4.49**	**2.84**	42.48	42.00
	(2.35)	(1.75)	(1.67)	(1.34)	[26.16]/1,649	[26.64]/2,106
Father in HH, age 0-3	0.006	-0.000	-0.018	0.002	0.759	0.714
	(0.027)	(0.021)	(0.016)	(0.014)	[0.402]/2,033	[0.427]/2,891
Grandmother in HH, age 0-3	0.010	-0.018	0.021	-0.014	0.110	0.175
	(0.661)	(0.016)	(0.138)	(0.010)	[0.267]/2,137	[0.316]/3,622
Maternal care, age 0-3	**-0.059**	-0.023	**-0.089**	**-0.052**	0.618	0.650
	(0.026)	(0.016)	(0.020)	(0.013)	[0.413]/2,114	[0.410]/3,702
Relative care, age 0-3	0.008	0.013	0.026	0.021	0.184	0.177
	(0.023)	(0.014)	(0.018)	(0.012)	[0.320]/2,114	[0.323]/3,702
Non-relative care, age 0-3	**0.051**	0.010	**0.063**	**0.031**	0.198	0.173
	(0.025)	(0.013)	(0.018)	(0.012)	[0.336]/2,114	[0.320]/3,702
Breastfed	0.053	-0.011	-0.008	0.002	0.484	0.412
	(0.029)	(0.017)	(0.022)	(0.014)	[0.500]/1,933	[0.492]/3,511
Regular doctor visits, age 0-3	-0.012	-0.018	0.012	-0.015	0.476	0.468
	(0.053)	(0.040)	(0.035)	(0.029)	[0.500]/1,918	[0.499]/2,360
Ever at dentist, age 0-3	-0.043	-0.037	-0.035	-0.016	0.227	0.228
	(0.049)	(0.037)	(0.039)	(0.032)	[0.419]/1,449	[0.420]/1,798
Weight change during pregnancy	-1.08	-0.138	-0.238	-0.780	31.24	30.66
	(1.17)	(0.692)	(0.857)	(.549)	[14.96]/1,873	[14.98]/3,337
Illness, age 0-1	0.025	0.004	0.013	-0.014	0.564	0.537
	(0.043)	(0.026)	(0.033)	(0.021)	[0.496]/1,907	[0.499]/3,401
Premature birth	0.031	-0.016	-0.033	-0.014	0.231	0.228
	(0.037)	(0.021)	(0.028)	(0.018)	[0.422]/1,902	[0.416]/3,400
Private health insurance, age 0-3	0.004	-0.015	-0.019	-0.014	0.697	0.655
	(0.033)	(0.026)	(0.018)	(0.017)	[0.422]/1,925	[0.446]/2,368
Medicaid, age 0-3	-0.038	0.008	-0.028	-0.015	0.265	0.267
	(0.026)	(0.024)	(0.015)	(0.014)	[0.411]/1,925	[0.417]/2,366
Log income, age 0-3	0.020	0.012	-0.051	-0.014	10.62	10.46
	(0.048)	(0.030)	(0.033)	(0.022)	[0.935]/2,052	[0.848]/3,546
Log income at age 3	0.044	0.018	0.036	0.019	10.59	10.43
	(0.068)	(0.045)	(0.050)	(0.034)	[1.05]/1,734	[0.971]/2,952
Mom avg. hours worked,	-2.24	-1.89	0.239	-0.004	30.23	28.79
year before birth	(1.35)	(1.12)	(0.970)	(0.773)	[14.10]/1,564	[13.54]/1,905
Mom avg. hours worked, age 0-1	-1.15	-0.453	0.763	1.30	34.05	32.95
	(1.23)	(1.16)	(0.963)	(0.752)	[12.77]/1,144	[12.26]/1,639
Mom smoked before birth	0.005	0.001	0.030	0.016	0.294	0.336
	(0.024)	(0.017)	(0.018)	(0.013)	[0.456]/1,917	[0.472]/3,430
Mom drank before birth	0.003	0.001	0.010	0.003	0.059	0.083
	(0.022)	(0.013)	(0.494)	(0.010)	[0.235]/2,144	[0.276]/3,738

Notes: Standard errors are in parenthesis. Estimates in bold were significant at the 5 percent level or less. Standard deviations are in brackets, followed by cohort sample size for that covariate. Each covariate was regressed on preschool status indicators (Head Start; other 'Preschool') with 'No preschool' as the reference status (i.e., the control mean).

Table S6. Selection Bias Estimates: Sibling Differences in Pre-Treatment Covariates by Preschool Status Across Cohorts

	Head Start estimate (se)				Preschool estimate (se)/*p*-value[b]				None: Control mean [sd]/Overall sample size		
	Deming's cohort	Complement cohort	Combined cohorts	*p*-value[a]	Deming's cohort	Complement cohort	Combined cohorts	*p*-value[a]	Deming replicated	Complement cohort	Combined cohorts
Pre-treatment index	.014 (.061)	-.001 (.063)	.005 (.035)	.098	.047 (.055)	.031 (.039)	.018 (.027)	.536	-.063 [.986]/1251	-.003 [.975]/2,144	-.031 [.983]/3,738
PPVT at 3	2.24 (4.82)	1.58 (3.38)	1.78 (1.99)	.774	-7.16* (4.12)/.070	-.535 (3.73)	-1.01 (2.05)	.740	19.90 [11.10]/195	22.00 [13.25]/396	21.05 [12.60]/644
Log of birth weight	.048** (.020)	-.008 (.018)	.023* (.012)	.031	-.006 (.017)/.064	-.025** (.012)	-.010 (.009)/.015	.506	4.70 [.25]/1,226	4.76 [.203]/1,894	4.73 [.225]/3,449
Very low BW	-.022* (.012)	.006 (.010)	-.007 (.007)	.095	-.004 (.008)	.010 (.007)	.005 (.004)	.227	.021 [.145]/1,226	.006 [.074]/1,894	.013 [.114]/3,449
In mother's HH, 0-3	.002 (.029)	.003 (.007)	.008 (.007)	.807	-.028 (.027)	-.002 (.005)	-.001 (.005)	.952	.900 [.302]/1,187	.988 [.085]/2,123	.974 [.119]/3,601
Previous health limitation	-.001 (.014)	.001 (.026)	.008 (.014)	.387	-.041** (.018)/.052	-.045** (.018)/.078	-.020* (.012)/.064	.750	.405 [.197]/1,187	.095 [.293]/2,123	.066 [.249]/3,601
Firstborn	.016 (.055)	.100* (.052)	-.016 (.031)	.024	-.124** (.055)/.038	.277*** (.039)/.0009	.051* (.026)/.046	<.001	.419 [.494]/1,251	.183 [.387]/2,144	.319 [.466]/3,738
Male	-.000 (.048)	-.019 (.043)	-.004 (.026)	.927	-.003 (.046)	-.054 (.115)	-.010 (.023)	.092	.501 [.500]/1,251	.528 [.499]/2,144	.512 [.500]/3,738
HOME at 3	1.98 (3.25)	.829 (2.35)	-.247 (1.75)	.693	3.03 (4.10)	4.49** (1.67)	2.84** (1.34)/.087	.350	38.05 [26.25]/427	42.48 [26.16]/1,649	42.00 [26.64]/2,106
Father in HH, 0-3	.009 (.034)	.006 (.027)	-.000 (.021)	.580	-.003 (.023)	-.018 (.016)	.002 (.014)	.747	.624 [.450]/739	.759 [.402]/2,033	.714 [.427]/2,891
Grandmother in HH, 0-3	-.003 (.024)	.010 (.661)	-.018 (.016)	.627	-.049*** (.019)/.078	.021 (.138)	-.014 (.010)	.024	.215 [.325]/1,190	.110 [.267]/2,137	.175 [.316]/3,622
Maternal care, 0-3	.0187 (.019)	-.059** (.026)	-.023 (.016)	.064	-.015 (.483)	-.089*** (.020)	-.052*** (.013)/.084	.097	.689 [.405]/1,244	.618 [.413]/2,114	.650 [.410]/3,702
Relative care, 0-3	-.007 (.019)	.008 (.023)	.013 (.014)	.232	.022 (.019)	.026 (.018)	.021* (.012)	.766	.180 [.335]/1,244	.184 [.320]/2,114	.177 [.323]/3,702
Non-relative care, 0-3	-.012 (.017)	.051** (.025)	.010 (.013)	.387	-.006 (.016)	.063*** (.018)	.031** (.012)	.020	.131 [.283]/1,244	.198 [.336]/2,114	.173 [.320]/3,702
Breastfed	-.053** (.027)	.053* (.029)	-.011 (.017)	.093	-.010 (.024)	-.008 (.022)/.042	.002 (.014)	.994	.333 [.472]/1,234	.484 [.500]/1,933	.412 [.492]/3,511
Regular doctor visits, 0-3	.043 (.102)	-.012 (.053)	-.018 (.040)	.865	-.055 (.110)	.012 (.035)	-.015 (.029)	.707	.383 [.488]/430	.476 [.500]/1,918	.468 [.499]/2,360

Table S6. Selection Bias Estimates: Sibling Differences in Pre-Treatment Covariates by Preschool Status Across Cohorts

	Head Start estimate (se)				Preschool estimate (se)/*p*-value[b]				None: Control mean [sd]/Overall sample size		
	Deming's cohort	Complement cohort	Combined cohorts	*p*-value[a]	Deming's cohort	Complement cohort	Combined cohorts	*p*-value[a]	Deming replicated	Complement cohort	Combined cohorts
Ever at dentist, 0-3	.033 (.137)	-.043 (.049)	-.037 (.037)	.586	.008 (.137)	-.035 (.039)	-.016 (.032)	.583	.303 [.461]/401	.227 [.419]/1,449	.228 [.420]/1,798
Weight change during pregnancy	.056 (1.18)	-1.08 (1.17)	-.138 (.692)	.840	-.168 (1.14)	-.238 (.857)	-.780 (.549)	.692	29.71 [15.34]/1,146	31.24 [14.96]/1,873	30.66 [14.98]/3,337
Illness, 0-1	.016 (.042)	.025 (.043)	.004 (.026)	.721	-.061 (.041)	.013 (.033)	-.014 (.021)	.123	.520 [.500]/1,175	.564 [.496]/1,907	.537 [.499]/3,401
Premature birth	-.047 (.034)	.031 (.037)	-.016 (.021)	.093	.007 (.034)	-.033 (.028)	-.014 (.018)	.255	.218 [.413]/1,175	.231 [.422]/1,902	.228 [.416]/3,400
Private health insurance, 0-3	.093 (.069)	.004 (.033)	-.015 (.026)	.096	.032 (.049)	-.019 (.018)	-.014 (.017)	.151	.447 [.481]/431	.697 [.422]/1,925	.655 [.446]/2,368
Medicaid, 0-3	.048 (.061)	-.038 (.026)	.008 (.024)	.984	-.006 (.043)	-.028* (.015)	-.015 (.014)	.526	.376 [.456]/431	.265 [.411]/1,925	.267 [.417]/2,366
Log income, 0-3	-.012 (.040)	.020 (.048)	.012 (.030)	.589	.040 (.033)	-.051 (.033)	-.014 (.022)	.624	10.00 [.718]/1,186	10.62 [.935]/2,052	10.46 [.848]/3,546
Log income, at 3	.011 (.085)	.044 (.068)	.018 (.045)	.721	.054 (.064)	.036 (.050)	.019 (.034)	.730	9.98 [.826]/993	10.59 [1.05]/1,734	10.43 [.971]/2,952
Mom avg. hours worked, year b. birth	-1.11 (3.14)	-2.24* (1.35)	-1.89* (1.12)	.500	2.06 (1.87)	.239 (.970)	-.004 (.773)/.093	.185	26.03 [12.15]/377	30.23 [14.10]/1,564	28.79 [13.54]/1,905
Mom avg. hours worked, 0-1	-1.08 (3.17)	-1.15 (1.23)	-.453 (1.16)	.797	1.77 (1.72)	.763 (.963)	1.30* (.752)	.356	32.52 [11.07]/379	34.05 [12.77]/1,144	32.95 [12.26]/1,639
Mom smoked b. birth	-.012 (.031)	.005 (.024)	.001 (.017)	.821	-.005 (.023)	.030* (.018)	.016 (.013)	.399	.392 [.489]/1,186	.294 [.456]/1,917	.336 [.472]/3,430
Mom drank b. birth	.004 (.021)	.003 (.022)	.001 (.013)	.577	.010 (.021)	.010 (.494)	.003 (.010)	.911	.081 [.272]/1,251	.059 [.235]/2,144	.083 [.276]/3,738

Notes. *** $p < .01$; ** $p < .05$; * $p < .1$. [a] *p*-value corresponds to estimates difference tests between Deming's cohort and complement cohort, both included in the Combined cohorts. [b] *p*-value <.1: estimate difference test between Head Start and Preschool status.

Table S7. Selection Bias Estimates: Sibling Differences in Attrition and Age by Preschool Status Across Cohorts

	Head Start estimate (se)	Preschool estimate (se)	None: Control mean [sd]/Overall sample size
Deming's cohort — Non-attrited[a] observations	.014 (.021)	.032 (.022)	.860 [.347]/1,384
Complement cohort — Attrited observations	-.029** (.013)	-.023** (.012)	.045 [.207]/2,215
Combined cohorts — Attrited observation	-.009 (.007)	-.011 (.007)	.025 [.169]/3,826
Deming's cohort — Age in 2004[b]	.182 (.298)	-.433 (.249)	23.20 [2.88]/1,251
Complement cohort — Age in 2014	1.07** (.425)	1.80*** (.270)	24.45 [3.95]/2,144
Combined cohorts — Age in 2014	.085 (.272)	.415** (.212)	29.03 [5.75]/3,738

Notes. *** $p < .01$; ** $p < .05$; * $p < .1$. [a] For the Deming's cohort, non-attrited observations count is considered instead of attrited due to different cross-round year of last interview variable available in NLSY, and used to derived attrition counts. Instead, it was possible to derive non-attrition count from by-survey-round dichotomous interview variables. It is non-attrition status that is used in FE sample eligibility: in Deming's cohort, the same non-attrition counts as in Deming (2009) was obtained. [b] Age by 2004 (i.e., at least 4 by 1990) in Deming's cohort is of course different when taken for the same cohort a decade later in the combined cohorts (i.e., at least 4 by 2000): therefore, these selection bias estimates are not readily comparable across cohorts unlike the other covariates and are considered separately.

Table S8. Head Start Impacts on Averaged Test & BPI Scores; Nontest & Adulthood Summary Index Scores, Across Cohorts (Overall & by Subgroups)

	Test (5-14)					BPI (5-14)				Nontest score index (7-14)					Adulthood index (19+)					
	Deming (2009)[a]	Repl.[b]	Comp.[c]	Comb.[d]	*p*[e]	Repl.	Comp.	Comb.	*p*	Deming (2009)	Repl.	Comp.	Comb.	*p*	Deming (2009)	Repl.	Ext.[f] (28+)	Comp.	Comb.	*p*
Overall																				
Head Start	.101	.106*	-.003	-.007	.243	.056	.069	.071**	.816	.265***	.262***	-.149*	-.001	.001	.228***	.202***	.166**	-.145**	-.011	.005
	(.057)	(.056)	(.062)	(.036)		(.047)	(.048)	(.029)		(.082)	(.081)	(.080)	(.047)		(.072)	(.072)	(.069)	(.068)	(.041)	
Preschool	-.012	.010	.015	.006	.682	.040	.109***	.074***	.234	.172*	.177**	-.124**	-.028	.008	.069	.091	.034	-.040	-.003	.644
	(.062)	(.060)	(.042)	(.030)		(.047)	(.040)	(.026)		(.088)	(.088)	(.060)	(.040)		(.072)	(.074)	(.066)	(.053)	(.035)	
p (HS=preschool)	.118	.175	.801	.731		.762	.369	.924		.372	.420	.747	.605		.080	.237	.134	.148	.859	
Sample size	1,251	1,251	2,144	3,738		1,251	2,144	3,738		1,251	1,251	2,141	3,734		1,251	1,251	1,251	2,144	3,738	
[HS/Preschool]	[364/364]	[364/364]	[497/795]	[951/1,275]		[364/364]	[497/795]	[951/1,275]		[364/364]	[364/364]	[497/795]	[951/1,275]		[364/364]	[364/364]	[364/364]	[497/795]	[951,1,275]	
By race																				
Head Start	.110	.110	-.097	-.054	.081	.094	.098	.112***	.762	.177	.173	-.155	-.005	.001	.237**	.177	.153	-.087	-.011	.154
(white/Hispanic)	(.090)	(.090)	(.091)	(.054)		(.063)	(.068)	(.040)		(.111)	(.110)	(.108)	(.068)		(.103)	(.100)	(.102)	(.102)	(.062)	
Head Start	.107	.108	.062	.041	.798	.028	.003	.031	.978	.351***	.348***	-135	.008	.036	.224**	.234**	.182**	-.180*	-.017	.012
(black)	(.072)	(.072)	(.083)	(.049)		(.067)	(.067)	(.042)		(.120)	(.119)	(.122)	(.066)		(.102)	(.106)	(.092)	(.099)	(.056)	
p (nonblack=black)	.982	.988	.191	.184		.476	.323	.162		.282	.277	.900	.896		.924	.697	.831	.531	.945	
Sample size	1,251	1,251	2,144	3,738		1,251	2,144	3,738		1,251	1,251	2,141	3,734		1,251	1,251	1,251	2,144	3,738	
[nonblack/black]	[695/556]	[695/556]	[1,375/769]	[2,280/1,458]		[695/556]	[1,375/769]	[2,280/1,458]		[695/556]	[695/556]	[1,374/767]	[2,278/1,456]		[695/556]	[695/556]	[695/556]	[1,375/769]	[2,280/1,458]	
By gender																				
Head Start	.159**	.160**	-.065	-.019	.118	.159**	.099	.098**	.953	.390***	.385***	-.223*	-.003	.005	.182*	.130	.103	-.195**	-.063	.058
(male)	(.076)	(.076)	(.085)	(.049)		(.067)	(.064)	(.040)		(.123)	(.122)	(.117)	(.069)		(.103)	(.098)	(.101)	(.100)	(.058)	
Head Start (female)	.055	.055	.056	.005	.742	-.035	.037	.044	.805	.146	.144	-.054	.002	.008	.272**	.272***	.226**	-.090	.042	.019
	(.081)	(.081)	(.077)	(.048)		(.064)	(.067)	(.041)		(.108)	(.108)	(.102)	(.063)		(.106)	(.96)	(.112)	(.093)	(.061)	
p (male=female)	.346	.343	.246	.716		.034	.489	.331		.135	.140	.262	.961		.533	.273	.450	.414	.220	
Sample size	1,251	1,251	2,144	3,738		1,251	2,144	3,738		1,251	1,251	2,141	3,734		1,251	1,251	1,251	2,144	3,738	
[male/female]	[627/624]	[627/624]	[1,104/1,040]	[1,904/1,834]		[627/624]	[1,104/1,040]	[1,904/1,834]		[627/624]	[627/624]	[1,102/1,039]	[1,901/1,833]		[627/624]	[627/624]	[627/624]	[1,104/1,040]	[1,904/1,834]	
By maternal AFQT score																				
Head Start	.015	-.006	.054	-.030	.083	-.065	.037	.017	.946	.529***	.509***	-.212	.083	.106	.279**	.317***	.384***	.027	.108	.147
(AFQT ≤ -1)	(.094)	(.094)	(.136)	(.067)		(.062)	(.101)	(.050)		(.156)	(.153)	(.193)	(.093)		(.114)	(.119)	(.115)	(.145)	(.080)	
Head Start	.154**	.168**	-.022	.004	.008	.124*	.079	.091**	.802	.124	.123	-.094	-.028	.003	.202**	.166*	.046	-.178**	-.053	.041
(AFQT > -1)	(.071)	(.071)	(.068)	(.043)		(.064)	(.053)	(.036)		(.091)	(.091)	(.087)	(.054)		(.091)	(.090)	(.084)	(.079)	(.048)	
p (low=high AFQT)	.245	.146	.617	.671		.034	.709	.227		.024	.030	.573	.303		.595	.309	.017	.218	.086	
Sample size	1,251	1,251	2,144	3,738		1,251	2,144	3,738		1,251	1,251	2,141	3,734		1,251	1,251	1,251	2,144	3,738	
[low/high]	[365/886]	[365/886]	[365/1,779]	[810/2,928]		[365/886]	[365/1,779]	[810/2,928]		[365/886]	[365/886]	[364/1,777]	[808/2,926]		[365/886]	[365/886]	[365/886]	[365/1,779]	[810/2,928]	

Notes. *** $p < .01$; ** $p < .05$; * $p < .1$. [a] Deming's (2009) published estimates. [b] Repl. = Deming (2009) replicated. [c] Comp. = Complement cohort includes siblings fitting the same criteria as in Deming (2009) but found eligible from 1990 to 2000. [d] Comb. = Combined cohorts includes siblings up to 2000 (i.e., integrate both Deming's and the complement cohorts). [e] *p*-value for estimates' difference testing between Ext. and Complement cohort. [f] Ext. = Deming's (2009) cohort (i.e., siblings eligible up to 1990), with outcomes extended to last available NLSY 2014 survey round (by then, siblings are all 28+ years old).

Table S9. Head Start Impacts on Individual Outcomes, Across Cohorts (Overall & by Subgroups)

	Grade retention				Learning disability diagnosis				Teen parenthood				High school graduation				
	Repl.[a]	Comp.[b]	Comb.[c]	*p* [d]	Repl.	Comp.	Comb.	*p*	Repl.	Comp.	Comb.	*p*	Repl.	Ext.[e] (28+)	Comp.	Comb.	*p*
Overall																	
Head Start	-.069*	.010	-.005	.137	-.059***	.042**	.003	.002	-.002	.034	-.002	.338	.087***	.020	-.000	.007	.905
	(.040)	(.034)	(.022)		.021	(.021)	(.012)		(.036)	(.032)	(.021)		(.032)	(.032)	(.029)	(.019)	
Preschool	-.086**	.022	.006	.003	-.023	.027*	.009	.286	-.057	.027	-.006	.330	-.006	-.025	.015	-.005	.218
	(.037)	(.023)	(.018)		(.031)	(.016)	(.010)		(.035)	(.020)	(.016)		(.030)	(.032)	(.020)	(.016)	
p (HS=preschool)	.680	.709	.603		.187	.480	.673		.229	.816	.853		.015	.262	.599	.549	
By race																	
Head Start	-.029	.029	-.012	.495	-.046	.034	.010	.002	.014	.030	-.010	.927	.046	.032	-.054	-.006	.578
(white/Hispanic)	(.059)	(.044)	(.031)		(.030)	(.029)	(.018)		(.053)	(.041)	(.028)		(.050)	(.050)	(.043)	(.029)	
Head Start	-.103*	-.015	-.002	.143	-.071**	.056*	-.004	.134	-.025	.061	.020	.164	.121***	.007	.066*	.018	.434
(black)	(.055)	(.052)	(.033)		(.028)	(.031)	(.015)		(.051)	(.050)	(.030)		(.041)	(.040)	(.040)	(.027)	
p (nonblack=black)	.357	.521	.816		.527	.594	.547		.593	.621	.463		.255	.695	.042	.535	
By gender																	
Head Start	-.203***	.056	-.031	.028	-.047	.044	.013	.043	.032	.025	-.004	.104	.119**	.036	-.029	.001	.799
(male)	(.058)	(.046)	(.033)		(.030)	(.030)	(.017)		(.053)	(.038)	(.026)		(.050)	(.050)	(.042)	(.029)	
Head Start (female)	.056	-.037	.020	.817	-.070***	.040*	-.007	.002	-.034	.041	.000	.919	.056	.004	.027	.014	.643
	(.056)	(.044)	(.030)		(.026)	(.023)	(.015)		(.056)	(.048)	(.031)		(.044)	(.044)	(.038)	(.026)	
p (male=female)	.002	.124	.243		.563	.913	.374		.415	.793	.922		.368	.646	.323	.745	
By maternal AFQT score																	
Head Start	-.128*	-.037	-.027	.833	-108***	.074	-.018	.099	-.033	-.023	-.023	.560	.163***	.059	.188***	.077*	.181
(AFQT ≤ -1)	(.067)	(.076)	(.043)		(.041)	(.048)	(.023)		(.064)	(.075)	(.041)		(.057)	(.053)	(.071)	(.039)	
Head Start	-.033	.013	.001	.143	-.031	.031	.011	.008	.016	.050	.008	.529	.044	-.004	-.044	-.017	.847
(AFQT > -1)	(.050)	(.037)	(.026)		(.022)	(.022)	(.013)		(.043)	(.035)	(.023)		(.037)	(.039)	(.031)	(.022)	
p (low=high AFQT)	.252	.543	.586		.095	.402	.266		.530	.370	.508		.080	.322	.003	.037	

Notes. *** $p < .01$; ** $p < .05$; * $p < .1$. [a] Repl. = Deming (2009) replicated. [b] Complement cohort includes siblings fitting the same criteria as in Deming (2009) but found eligible from 1990 to 2000. [c] Combined cohorts includes siblings up to 2000 (i.e., integrate both Deming's and the complement cohorts). [d] *p*-value for estimates' difference testing between Ext. and Complement cohort. [e] Ext. = Deming's (2009) cohort (i.e., siblings eligible up to 1990), with outcomes extended to last available NLSY 2014 survey round (by then, siblings are all 28+ years old).

Table S10. Head Start Impacts on Individual Outcomes, Across Cohorts (Overall & by Subgroups)

	Some college attended					Idle					Crime					Poor health status				
	Repl.[a]	Ext.[b] (28+).	Comp.[c]	Comb.[d]	*p* [e]	Repl.	Ext. (28+)	Comp.	Comb.	*p*	Repl.	Ext. (28+)	Comp.	Comb.	*p*	Repl.	Ext. (28+)	Comp.	Comb.	*p*
Overall																				
Head Start	.022	.110***	-.074*	-.002	.010	-.059*	.002	.082***	.016	.026	-.002	-.028	.033	.009	.071	-.069**	-.047	-.025	-.003	.459
	(.035)	(.039)	(.039)	(.023)		(.035)	(.033)	(.030)	(.020)		(.038)	(.041)	(.031)	(.021)		(.027)	(.032)	(.031)	(.019)	
Preschool	.061*	.053	-.043	.010	.030	.005	.019	.028	.005	.458	.001	-.001	.014	.012	.905	-.020	-.009	-.033	-.003	.253
	(.036)	(.042)	(.030)	(.020)		(.029)	(.031)	(.023)	(.016)		(.037)	(.036)	(.026)	(.018)		(.032)	(.031)	(.021)	(.015)	
p (HS=preschool)	.379	.269	.440	.627		.114	.648	.074	.566		.933	.571	.569	.892		.153	.327	.796	.993	
By race																				
Head Start	-.046	.108	-.057	-.011	.071	-.091*	-.012	.070	.027	.047	-.018	-.044	-.016	-.007	.417	-.092**	-.038	-.074	-.010	.315
(white/Hispanic)	(.048)	(.054)	(.053)	(.031)		(.053)	(.047)	(.045)	(.028)		(.058)	(.063)	(.043)	(.032)		(.043)	(.054)	(.045)	(.029)	
Head Start	.083*	.118**	-.093*	.004	.037	-.036	.014	.079*	.000	.242	.010	-.014	.057	.015	.086	-.047	-.049	.020	.009	.040
(black)	(.050)	(.055)	(.057)	(.034)		(.045)	(.049)	(.041)	(.028)		(.050)	(.052)	(.044)	(.029)		(.035)	(.037)	(.045)	(.025)	
p (nonblack=black)	.061	.891	.643	.741		.424	.707	.873	.502		.710	.704	.240	.601		.425	.865	.155	.610	
By gender																				
Head Start	-.029	.075	-.130**	-.024	.063	-.088*	-.025	.111***	.055**	.162	.025	-.043	.101**	.042	.331	-.035	-.020	-.107***	-.020	.612
(male)	(.046)	(.054)	(.054)	(.031)		(.045)	(.046)	(.042)	(.025)		(.058)	(.060)	(.045)	(.031)		(.037)	(.046)	(.038)	(.023)	
Head Start (female)	.070	.144***	-.016	.019	.028	-.032	.028	.049	-.023	.056	-.029	-.014	-.037	-.024	.073	-.103**	-.072	.057	.014	.109
	(.050)	(.053)	(.054)	(.033)		(.048)	(.052)	(.044)	(.030)		(.053)	(.057)	(.043)	(.029)		(.042)	(.049)	(.042)	(.028)	
p (male=female)	.135	.350	.130	.342		.376	.464	.317	.045		.505	.740	.027	.128		.247	.461	.002	.329	
By maternal AFQT score																				
Head Start	.001	.150***	-.089	.022	.161	-.064	-.078	.153**	.039	.035	-.009	.011	.050	.024	.466	-.091**	-.157***	-.074	-.078**	.299
(AFQT ≤ -1)	(.049)	(.056)	(.078)	(.040)		(.064)	(.062)	(.064)	(.041)		(.068)	(.072)	(.062)	(.041)		(.046)	(.059)	(.067)	(.037)	
Head Start	.034	.094*	-.066	-.009	.025	-.057	.045	.065*	.008	.142	.001	-.051	.011	.000	.107	-.058*	.014	-.007	.027	.903
(AFQT > -1)	(.047)	(.052)	(.045)	(.028)		(.042)	(.039)	(.033)	(.022)		(.045)	(.049)	(.036)	(.025)		(.034)	(.037)	(.035)	(.022)	
p (low=high AFQT)	.621	.473	.800	.529		.924	.092	.222	.508		.905	.477	.587	.626		.572	.015	.371	.014	

Notes. *** $p < .01$; ** $p < .05$; * $p < .1$. [a] Repl. = Deming (2009) replicated. [b] Ext. = Deming's (2009) cohort (i.e., siblings eligible up to 1990), with outcomes extended to last available NLSY 2014 survey round (by then, siblings are all 28+ years old). [c] Comp. = Complement cohort includes siblings fitting the same criteria as in Deming (2009) but found eligible from 1990 to 2000. [d] Comb. = Combined cohorts includes siblings up to 2000 (i.e., integrate both Deming's and the complement cohorts). [e] *p*-value for estimates' difference testing between Deming's cohort and Complement cohort.

Table S11. Head Start/Preschool Longer-Run Impacts on Adulthood Index Score, Educational Outcomes & Earnings (Overall and by Subgroups)

	Adulthood index		Educational attainment	College graduation	Earnings
	Deming (2009)	Deming's cohort (extended outcomes)[a]	Deming's cohort (extended outcomes)	Deming's cohort (extended outcomes)	Deming's cohort (extended outcomes)
Overall					
Head Start	.228***	.166**	.306**	-.016	.067
	(.072)	(.069)	(.145)	(.028)	(.122)
Preschool	.069	.034	.138	.006	-.049
	(.072)	(.066)	(.169)	(.028)	(.114)
p (HS=preschool)	.080	.134	.410	.452	.429
Sample size	1,251	1,251	1,251	1,251	1,195
[HS/Preschool]	[364/364]	[364/364]	[364/364]	[364/364]	[350/343]
By race					
Head Start	.224**	.153	.269	-.051*	.191
(white/Hispanic)	(.102)	(.102)	(.203)	(.029)	(.193)
Head Start	.237**	.182**	.334	.003	-.037
(black)	(.103)	(.092)	(.207)	(.032)	(.165)
p (nonblack=black)	.924	.831	.822	.218	.373
Sample size	1,251	1,251	1,251	1,251	1,195
[nonblack/black]	[695/556]	[695/556]	[695/556]	[695/556]	[656/539]
By gender					
Head Start	.182*	.103	.268	-.004	-.026
(male)	(.103)	(.101)	(.206)	(.032)	(.174)
Head Start	.272**	.226**	.343*	-.028	.153
(female)	(.106)	(.112)	(.205)	(.033)	(.186)
p (male=female)	.553	.450	.797	.627	.501
Sample size	1,251	1,251	1,251	1,251	1,195
[male/female]	[627/624]	[627/624]	[627/624]	[627/624]	[581/614]
By maternal AFQT score					
Head Start	.279**	.384***	.454**	-.005	.437**
(AFQT ≤ -1)	(.114)	(.115)	(.227)	(.031)	(.197)
Head Start	.202**	.046	.238	-.022	-.132
(AFQT > -1)	(.091)	(.084)	(.187)	(.029)	(.153)
p (low=high AFQT)	.595	.017	.462	.688	.023
Sample size	1,251	1,251	1,251	1,251	1,195
[low/high]	[365/886]	[365/886]	[365/886]	[365/886]	[354/841]

Notes. [a] Deming's cohort (i.e., siblings eligible up to 1990), with outcomes extended to last available NLSY 2014 survey round (by then, siblings are all 28+ years old). Adulthood index (in standard deviation) = composite of 6 indicators for high school graduation; college attendance; teen-age parenthood; 'idle'; involvement with the justice system; and poor health status. Educational attainment = amount of years of completed schooling. College graduation indicator (1= 16+ years of completed schooling); Earnings = natural log of averaged yearly earnings, in 2014 dollars, adjusted for age and survey-round year. Pretreatment covariates and sibling fixed effects are included throughout. 'No preschool' status is the counterfactual. Standard errors, in parenthesis, are clustered at the family level. ******* $p < .01$; ****** $p < .05$; ***** $p < .1$.

Table S12. Head Start & Preschool Impacts on Cognitive Test Scores Across Cohorts (by age groups)

	(1)					(2)					(3)					(4)					(5)				
	Deming (2009)[a]	Rep.[b]	Comp.[c]	Comb.[d]	*p* [e]	Deming (2009)	Rep.	Comp.	Comb.	*p*	Deming (2009)	Rep.	Comp.	Comb.	*p*	Deming (2009)	Rep.	Comp.	Comb.	*p*	Deming (2009)	Rep.	Comp.	Comb.	*p*
Head Start																									
5-6	-.025	-.027	-.249***	-.159***	.074	.081	.080	-.016	.043	.536	.093	.103	-.001	.054	.470	.131	.129	.055	.053	.122	.145*	.145*	.062	.041	.239
	(.091)	(.091)	(.065)	(.052)		(.083)	(.083)	(.063)	(.049)		(.079)	(.079)	(.061)	(.048)		(.087)	(.087)	(.070)	(.048)		(.085)	(.085)	(.069)	(.048)	
7-10	-.116	-.117	-.334***	-.233***	.038	.040	.039	-.108*	-.026	.299	.067	.078	-.090*	-.008	.185	.116*	.116*	-.037	-.005	.067	.133**	.133**	-.030	-.013	.115
	(.072)	(.072)	(.059)	(.046)		(.065)	(.065)	(.057)	(.042)		(.061)	(.060)	(.053)	(.039)		(.060)	(.060)	(.063)	(.039)		(.060)	(.060)	(.062)	(.040)	
11-14	-.201***	-.201***	-.319***	-.233***	.088	-.053	-.053	-.101*	-.039	.486	-.017	-.009	-.083	-.020	.395	.029	.029	-.053	-.028	.231	.055	.056	-.048	-.035	.255
	(.070)	(.070)	(.063)	(.045)		(.065)	(.065)	(.059)	(.041)		(.061)	(.060)	(.056)	(.039)		(.061)	(.061)	(.067)	(.040)		(.062)	(.062)	(.066)	(.040)	
Preschool																									
5-6	.167**	.161*	.264***	.233***	.211	.022	.014	.112**	.085**	.442	-.019	-.024	.094*	.053	.285	-.102	-.107	.095*	.031	.098	-.079	-.081	.079	.017	.109
	(.083)	(.083)	(.052)	(.043)		(.082)	(.081)	(.049)	(.040)		(.078)	(.077)	(.048)	(.039)		(.084)	(.084)	(.051)	(.038)		(.085)	(.085)	(.051)	(.039)	
7-10	.230***	.228***	.232***	.237***	.955	.111*	.108*	.067	.094***	.243	.087	.087	.053	.068**	.290	.031	.031	.030	.034	.960	.048	.049	.015	.024	.875
	(.070)	(.70)	(.049)	(.037)		(.064)	(.064)	(.044)	(.033)		(.061)	(.061)	(.042)	(.032)		(.061)	(.061)	(.045)	(.031)		(.065)	(.065)	(.046)	(.032)	
11-14	.182**	.182**	.204***	.189***	.768	.076	.074	.032	.056	.133	.037	.039	.004	.019	.148	-.040	-.039	-.032	-.015	.865	-.022	-.020	-.040	-.023	.856
	(.072)	(.072)	(.055)	(.040)		(.068)	(.068)	(.048)	(.036)		(.065)	(.064)	(.046)	(.034)		(.066)	(.066)	(.049)	(.035)		(.069)	(.069)	(.049)	(.035)	
Permanent income											.112*	.118 **	.102***	.100***											
											(.064)	(.055)	(.033)	(.028)											
Maternal AFQT											.353***	.345***	.259***	.241***											
											(.057)	(.057)	(.028)	(.026)											
Mom HS graduate											.141**	.238***	.241***	.272***											
											(.071)	(.074)	(.065)	(.050)											
Mom some coll.											.280***	.395***	.318***	.411***											
											(.080)	(.088)	(.072)	(.055)											
p all age Head Start effects =	.074	.077	.323	.231		.096	.100	.269	.226		.161	.151	.279	.290		.092	.097	.187	.217		.151	.155	.185	.257	
Baseline covariates	N	N	N	N		Y	Y	Y	Y		Y	Y	Y	Y		N	N	N	N		Y	Y	Y	Y	
Sibling FE	N	N	N	N		N	N	N	N		N	N	N	N		Y	Y	Y	Y		Y	Y	Y	Y	
Total # of scores	4,687	4,687	8,220	14,086		4,687	4,687	8,220	14,086		4,687	4,687	8,220	14,086		4,687	4,687	8,220	14,086		4,687	4,687	8,220	14,086	
R^2	.028	.028	.056	.053		.194	.195	.232	.236		.268	.273	.293	.295		.608	.607	.633	.574		.619	.618	.644	.582	
Sample size	1,251	1,251	2,144	3,738		1,251	1,251	2,144	3,738		1,251	1,251	2,144	3,738		1,251	1,251	2,144	3,738		1,251	1,251	2,144	3,738	

Notes: Test scores are standardized (mean = 0; SD = 1). *** $p < .01$; ** $p < .05$; * $p < .1$. [a] Deming's (2009) published estimates. [b] Repl. = Deming (2009) replicated. [c] Comp. = Complement cohort includes siblings fitting the same criteria as in Deming (2009) but found eligible from 1990 to 2000. [d] Comb. = Combined cohorts includes siblings up to 2000 (i.e., integrate both Deming's and the complement cohorts). [e] *p*-value for estimates' difference testing between Deming (2009) replicated. and Complement cohort. For further details on models (1) to (5) specifications refer to Deming (2009).

Table S13. Head Start & Preschool Impacts on BPI Scores Across Cohorts (by age groups)

	(1)				(2)				(3)				(4)				(5)			
	Repl.[a]	Comp.[b]	Comb.[c]	*p* [d]	Repl.	Comp.	Comb.	*p*	Repl.	Comp.	Comb.	*p*	Repl.	Comp.	Comb.	*p*	Repl.	Comp.	Comb.	*p*
Head Start																				
5-6	.122	.74	.140***	.587	.097	-.036	.058	.520	.080	-.022	.063	.552	.046	.024	.030	.906	.063	.060	.031	.878
	(.094)	(.063)	(.052)		(.314)	(.063)	(.051)		(.094)	(.063)	(.050)		(.085)	(.062)	(.045)		(.083)	(.061)	(.045)	
7-10	.149**	.134**	.194***	.509	.102	.017	.106**	.831	.084	.030	.108***	.679	.044	.079	.085**	.215	.053	.064	.085**	.274
	(.066)	(.055)	(.042)		(.067)	(.055)	(.041)		(.066)	(.055)	(.040)		(.057)	(.050)	(.033)		(.055)	(.050)	(.033)	
11-14	.098	.126**	.139***	.870	.062	.025	.061	.337	.042	.038	.064	.435	.054	.067	.049	.764	.063	.052	.048	.500
	(.068)	(.061)	(.044)		(.068)	(.061)	(.043)		(.067)	(.060)	(.043)		(.063)	(.058)	(.038)		(.063)	(.058)	(.038)	
Preschool																				
5-6	.074	-.103**	-.060	.023	.113	-.078	-.038	.089	.117	-.065	-.022	.068	.047	.053	.043	.951	.082	.052	.045	.762
	(.083)	(.052)	(.042)		(.082)	(.050)	(.042)		(.082)	(.051)	(.042)		(.073)	(.048)	(.037)		(.071)	(.050)	(.038)	
7-10	.099	-.054	-.021	.141	.117	-.030	.000	.259	.123	-.024	.012	.220	.054	.090**	.075**	.310	.077	.089**	.076**	.284
	(.064)	(.046)	(.035)		(.064)	(.045)	(.035)		(.063)	(.045)	(.035)		(.055)	(.042)	(.030)		(.055)	(.044)	(.031)	
11-14	-.044	.002	-.013	.923	-.022	.040	.005	.821	-.020	.044	.019	.782	-.003	.132***	.078**	.229	.023	.131***	.074**	.321
	(.067)	(.049)	(.037)		(.066)	(.048)	(.037)		(.066)	(.048)	(.037)		(.057)	(.046)	(.031)		(.058)	(.048)	(.032)	
Permanent income									-.175***	-.089***	-.133***									
									(.061)	(.034)	(.031)									
Maternal AFQT									.082	.080***	.086***									
									(.055)	(.030)	(.027)									
Mom HS graduate									-.242***	-.153*	-.238***									
									(.087)	(.096)	(.061)									
Mom some coll.									-.294***	-.221**	-.307***									
									(.096)	(.088)	(.067)									
p all age Head Start effects =	.77	.540	.264		.849	.572	.390		.839	.583	.410		.990	.536	.360		.987	.514	.350	
Baseline covariates	N	N	N		Y	Y	Y		Y	Y	Y		N	N	N		Y	Y	Y	
Sibling FE	N	N	N		N	N	N		N	N	N		Y	Y	Y		Y	Y	Y	
Total # of scores	4,610	8,303	14,098		4,610	8,303	14,098		4,610	8,303	14,098		4,610	8,303	14,098		4,610	8,303	14,098	
R^2	.025	.049	.060		.065	.108	.101		.080	.115	.114		.540	.579	.532		.553	.584	.534	
Sample size	1,251	2,144	3,738		1,251	2,144	3,738		1,251	2,144	3,738		1,251	2,144	3,738		1,251	2,144	3,738	

Notes: BPI scores are standardized (mean = 0; SD = 1). *** $p < .01$; ** $p < .05$; * $p < .1$. [a] Repl. = Deming (2009) replicated. [b] Comp. = Complement cohort includes siblings fitting the same criteria as in Deming (2009) but found eligible from 1990 to 2000. [c] Comb. = Combined cohorts includes siblings up to 2000 (i.e., integrate both Deming's and the complement cohorts). [d] *p*-value for estimates' difference testing between Deming (2009) replicated. and Complement cohort. For further details on models (1) to (5) specifications refer to Deming (2009).

Table S14. Head Start (HS) by cohort effect on adulthood summary index (measured in 2014), controlling for Head Start by covariate interaction(s)

	HS x Cohort	HS x Covariate	HS x Covariate (all interactions)
Main effect:	-.263 (.093)		-.293 (.164)
Including			
Head Start x *covariates:*			
Pre-treatment index	-.255 (.098)	-.016 (.155)	-.022 (.062)
Family human capital index	-.244 (.096)	-.057 (.057)	.002 (.071)
Gender	-.258 (.093)	-.123 (.098)	-.111 (.265)
Black	-.262 (.093)	.040 (.092)	.022 (.821)
Low maternal AFQT	-.236 (.093)	.210 (.104)	.227 (.119)
Mother's age at child's birth	-.228 (.137)	-.005 (.014)	-.006 (.023)
Mother's age at first child	-.272 (.107)	.006 (.013)	.022 (.018)
Mother's age at median child	-.275 (.118)	.004 (.014)	-.024 (.023)
Mother's age at last child	-.316 (.104)	.014 (.009)	.023 (.014)
Age at measurement	-.232 (.158)	.003 (.015)	-.018 (.025)

Notes. Family clustered standard error are in parenthesis. Base model corresponds to Equation (1) of the text. Thus, sibling fixed-effects and pretreatment covariates are included throughout. Pre-treatment index: standardized sum of all standardized pre-treatment covariates. Family human capital: standardized sum of standardized maternal and grandparent's years of completed schooling; maternal AFQT score, family permanent income (ln), and CNLSY H.O.M.E. scale score. Low maternal AFQT standardized score (mean of 0; sd of 1): ≤ -1. Mother's age at child's birth is a within-family variable. Mother's age at first, median, last child are between-families variables. Sample size = 3,141.