Educational Evaluation and Policy Analysis
Month 201X, Vol. 42, No. 3, pp. 1–22
DOI: 10.3102/0162373720948884
Article reuse guidelines: sagepub.com/journals-permissions
© 2020 AERA. http://eepa.aera.net

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

Remy Pages
University of California, Irvine
Dylan J. Lukes
Harvard Graduate School of Education
Drew H. Bailey
Greg J. Duncan
University of California, Irvine

Using an additional decade of data from the National Longitudinal Survey of Youth 1979 Children and Young Adults (CNLSY), this study replicated and extended Deming's 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 showed generally null impacts on school-age and early adulthood outcomes.

Keywords: *early childhood, evaluation, human development, longitudinal studies, regression analyses*

Introduction

Understanding the causal effects of early childhood programs implemented at scale on long-term adult outcomes is challenging. However, as early childhood is considered by many economists to be a key launching period for lifelong human capital accumulation (e.g., Chetty et al., 2011; Cunha et al., 2006; Currie & Almond, 2011; Heckman & Mosso, 2014; Hoynes et al., 2016), considerable attention has been devoted to research attempting to estimate the short- and longer-run impacts of early education programs (Duncan & Magnuson, 2013).

One such early childhood education program is Head Start, the United States' oldest and largest early childhood education program to be offered at scale. Given this, it is hardly surprising a great deal of the research has been devoted to it.[1] Studies of the longer-run impacts of Head Start attendance have generally shown positive, albeit sometimes mixed, results (Bauer & Schanzenbach, 2016; Carneiro & Ginja, 2014; Deming, 2009; Garces et al., 2002; Ludwig & Miller, 2007; Thompson, 2018). In a recent study analyzing data from the Panel Study of Income Dynamics (PSID) survey, Johnson and Jackson (2019) suggest that some of these inconsistencies can be attributed to complementarities between Head Start attendance and subsequent K–12 spending.[2]

One important study on the long-term impacts of Head Start attendance is Deming (2009). Using data from the National Longitudinal Survey of Youth 1979 Children and Young Adults (CNLSY), his paper builds on the approach of

Currie and Thomas (1995; also, Garces et al., 2002) by comparing both school-age and young adulthood outcomes between children who attended Head Start and their siblings who either attended other non–Head Start preschools or did not attend any preschool program. Most of the cohorts analyzed in Deming (2009) were born between 1976 and 1986 and had outcomes tracked through the survey's 2004 interviewing wave.[3] The study found that, compared with siblings who did not attend any preschool, children who attended Head Start averaged 8.5 percentage points (*pp*s) higher rates of high school graduation and 0.23 standard deviation (*SD*) higher scores on an index of adult outcomes.[4] Deming (2009) is noteworthy both for its sibling comparison 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.

However, recent research calls into question the sibling comparison design both in terms of its external validity (D. L. Miller et al., 2019) and potential to produce biased estimates from sibling spillover effects (Heckman & Karapakula, 2019). To correct for nonrandom selection into the family fixed effects (FFEs) identifying sample, D. L. Miller et al. (2019) found that reweighting on observables attenuated many of the original Deming (2009) FFE estimates of Head Start's impact on long-term outcomes. D. L. Miller et al. (2019) document similar attenuated FFE estimates of Head Start's long-term impact in the PSID. Highlighting threats to construct validity, Heckman and Karapakula (2019) found siblings who participated in the Perry Preschool Project had large positive spillovers on their nonparticipating siblings. This was particularly true for male siblings.

Importantly, evidence of Head Start's lasting positive impact into adulthood is not limited to Deming (2009) or the FFE design. Over the past several decades, a sizable body of evidence that leverages a variety of empirical methods, including FFE, regression discontinuity (RD), and difference-in-difference (DID), have shown Head Start's ability to improve adolescent and longer-term outcomes (Bailey et al., 2018; Barr & Gibbs, 2019; Bauer & Schanzenbach, 2016; Carneiro & Ginja, 2014; Currie & Thomas, 1995, 1999; Garces et al., 2002; Ludwig & Miller, 2007; D. L. Miller et al., 2019; Thompson, 2018). Most consistent across this body of research is Head Start's positive impact on educational attainment, health outcomes, and reduced criminal activity with estimated impacts tending to be larger and more robust for males, siblings from earlier birth cohorts, and those born to mothers with less than a high school education.

On the whole, these studies predominantly focused on 1970s and 1980s birth cohorts. Notable exceptions include Carneiro and Ginja's (2014) RD analysis of 1977–1996 birth cohorts and Bauer and Schanzenbach's (2016) FFE analysis of 1970–1990 birth cohorts. Although not part of their main results, Barr and Gibbs's (2019) supplementary FFE analysis sample (contained in their appendix) included CNLSY 1970–1992 birth cohorts. Results are mixed on Head Start's impact for more recent birth cohorts. Carneiro and Ginja (2014) indicated that a robustness check showed that the positive effects for males age 12 to 13 in their overall sample were driven by the earlier 1980s birth cohorts. Similarly, in an appendix FFE analysis, Barr and Gibbs (2019) found no significant impact of Head Start on high school graduation, some college, crime, teen parenthood, or their index of adulthood outcomes. However, Head Start impacts were positive and significant for males on high school graduation, crime, and their index of adulthood outcomes (Barr & Gibbs, 2019). In both cases, each overall sample included birth cohorts from 1970s through the early to mid-1990s. In contrast, Bauer and Schanzenbach's (2016) FFE analysis found positive impacts of Head Start on high school graduation, some higher education, postsecondary completion, self-control index, self-esteem index, and positive parenting index. These results more closely followed Deming (2009) and included birth cohorts up to 1990.[5] A detailed synthesis of these studies and more—including birth cohorts analyzed, identification strategies, and findings—can be found in the Supplementary Table A1 in the online version of the journal.

Present Study

The present work builds upon this rich existing literature by expanding Deming's (2009) evaluation of Head Start's longer-run impacts. By appending 10 additional years to the original 1976–1986 birth cohorts analyzed in Deming (2009), we were able to estimate impacts on outcomes measured later in adulthood and not previously considered: educational attainment, college graduation, and earnings. Second, the additional data provided us with an opportunity to apply the methods used in Deming (2009) to 10 additional birth cohorts in the National Longitudinal Survey of Youth (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 summary index (ASI) of Head Start attendance relative to not attending any preschool program from 0.23 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). This is notable and, taken by itself, could indicate a sizable return on investment for the program. However, we estimated relatively small, nonsignificant impacts on gains on other later life outcomes, including college graduation and earnings.[6] For the children born after Deming's cohorts, Head Start impacts were mostly null and sometimes negative. In fact, positive impacts on ASI generated by Deming's cohorts were matched by nearly symmetric negative impacts for the complement cohorts (−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 ASI was close to zero and not statistically significant.

In light of recent work by D. L. Miller et al. (2019), following these initial analyses, we checked whether our FFE identifying samples exhibited "selection into identification" (SI) across a variety of observable characteristics, including family size and mother's age at child's birth. Finding evidence of SI in both the Deming and combined cohorts, we used the one-step reweighting-on-observables procedure outlined in D. L. Miller et al. (2019) to correct for any potential bias. Similar to D. L. Miller et al. (2019), after reweighting the Deming cohort, we found attenuated estimates of Head Start's impact on long-term outcomes. However, for the combined cohorts sample, we found limited evidence that reweighting attenuated Head Start impact estimates on long-term outcomes.

Our article concludes with a discussion of what is driving these cross-cohort differences in Head Start's impact. Although we found differences in baseline human capital between cohorts, we found no evidence that the impact of Head Start varied for different levels of human capital within cohorts. Similarly, we observed differences in other pretreatment covariates between cohorts, but also found limited evidence that they drove variation in Head Start impacts. Finally, to better understand whether the effects of Head Start were changing across cohorts, we performed a Blinder–Oaxaca decomposition (Blinder, 1973; Oaxaca, 1973) using a threefold decomposition (Jann, 2008). In line with the above analyses, the results from this exercise also highlighted key differences between the Deming and complement cohort samples, but importantly showed how these differences—most notably in the pretreatment index and mother's age at child's birth—were associated with variation in estimated Head Start impacts across cohorts. These results indicated that if mother's age at child's birth was set fixed at Deming's cohort level, the ASI mean for the complement cohort's Head Start attendees would be similar to that of Deming's cohort counterparts.

These final analyses shed light on the potential importance non-Head Start factors play in moderating the direction and magnitude of Head Start impacts—most notably the pretreatment index and mother's age at child's birth. The pretreatment index is composed of a wide range of within-family covariates that vary between siblings and occur before or at the age of a child's Head Start eligibility. These variables, many of which have been shown to improve with each marginal year a young mother delays childbirth beyond her teenage years (e.g., Augustine et al., 2015; Duncan et al., 2018; Hotz et al., 2005; A. R. Miller, 2011), include maternal work history, maternal income, maternal and infant health, child care arrangements, and household structures. Holding all else equal, changes in these underlying measures could alter a given parent's need of and benefit from Head Start's wraparound services which both educate and, in some cases, employ parents of participating children (Currie & Neidell, 2007; Zigler & Valentine, 1979). Beyond this, children from better resourced households participating in Head Start may also need and benefit less from participation, particularly if household resources such as parental human capital, parental income, or parental socioemotional skills—each of which typically improve over a parent's lifetime—substitute for the benefits that otherwise would have been provided by the program.

Although we are unable to conduct definitive tests of them, these hypotheses are consistent with extant literature documenting heterogeneity in the impacts of Head Start on short- and long-term outcomes, with impacts, concentrated at the left tail of the outcome distribution suggesting Head Start benefits participants most in need, particularly children of parents with low skill or low social backgrounds (Bitler et al., 2014; De Haan & Leuven, 2020). Although results did not show robust evidence of heterogeneity of Head Start estimates within cohorts across time, our analyses hint at the disproportionate positive impact Head Start has for those children from less advantaged households relative to their more advantaged participating peers. Still, additional factors could help explain cross-cohort differences in program effects, including shifting Head Start counterfactuals, changes in Head Start programming and quality of offerings, and an altered social context. Although our article provides no definitive evidence regarding these claims, they could be promising avenues of future exploration.

Thus, although the past several years have seen a resurgence of research on the long-run impacts of Head Start, this study adds value in several notable ways. First, we used a well-established FFE design, which hitherto has estimated positive long-run outcomes of Head Start, to estimate predominantly negative or null long-run outcomes of the program and showed how, if at all, these results were sensitive to "selection into identification" by performing the one-step reweighting-on-observables procedure as outlined in D. L. Miller et al. (2019). Second, despite major changes to the program and social context of Head Start–eligible children during this period, ours is the first paper to estimate the impact of Head Start for the most recent set of CNLSY birth cohorts and to compare their program effects with those of earlier cohorts.[7] Finally, we attempted to reconcile why Head Start impacts were different across cohorts, finding suggestive evidence that between-cohort differences in the ages of mothers at the time of their child's birth played an important explanatory role.

Method

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% of whom were between the ages of 3 and 5, at an annual cost of around US$9 billion in federal funding (U.S. Department of Health and Human Services, Administration for Children and Families, 2018). Enrollment and funding have varied greatly since Head Start's 1965 inception. Participation grew until the early 1980s, plateaued through the early 1990s, and then grew again when funded enrollments almost doubled (i.e., from around 500,000 in 1990 to 900,000 in 2000). Appropriations (in 2018 US$) grew from about US$3 billion in 1990 to US$9 billion in 2000. After 2000, both enrollment and inflation-adjusted funding remained steady (U.S. Department of Health and Human Services, Administration for Children and Families, 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), enrollment increased by about 60%. However, the proportion of teachers or assistant teachers with at least a Child Development Associate credential increased very little—by about 5 percentage points (*pp*s) over this period (U.S. Department of Health and Human Services, Administration for Children and Families, 2018). More generally, the 1990s and 2000s were a time of rapid increases in preschool enrollment, including Head Start, but also state-run prekindergarten programs (Duncan & Magnuson, 2013).

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 effect (FE) analyses were estimated for a sample of siblings discordant on Head Start attendance and who enrolled in Head Start no later than 1990. Deming's sample eligibility rules were (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.

For our complement and combined cohorts, Deming's sample restrictions were moved forward by 10 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.[8] It is important to note that the sampling design of the CNLSY (i.e., all children were born to women who were between ages 14 and 22 in 1979) led children in our complement cohort to be born to older mothers than is the case for children in Deming's cohort. Later, we consider the role this factor may have played in explaining differences in Head's Start impact between the Deming and complement cohorts. Furthermore, because we wanted to both estimate Head Start's impacts on educational attainment, college graduation and earnings and assess their robustness on ASI, we both replicated and extended Deming's analysis for these cohorts up to 2014, the latest CNLSY survey round year available to us at the time of our analyses.

Family Background Statistics

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, and grandmother's highest grade completed.[9] Across these variables and for all three cohorts, there was a clear pattern of selection of more disadvantaged children 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 FE subsamples). Discrepancies between the two samples were small, suggesting that the demographic characteristics of the FE subsamples were similar to the less restricted, larger samples.[10]

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 Armed Forced Qualification Test (AFQT) than children not attending any preschool. Overall, Head Start children came from relatively more disadvantaged households.[11] 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.

Outcomes

As part of our replication and extension of Deming (2009), we assessed the impact of Head Start on the same set of three test scores, two nontest outcomes, and six young adulthood outcomes for each of the Deming, complement, and combined cohorts. The three test scores covered ages 5 to 14 and included the Peabody Picture Vocabulary Test (PPVT), the Peabody Individual Achievement Math (PIATMT) subtest, and the PIAT Reading Recognition (PIATRR) subtest. Following Deming (2009), due to the biannual survey design of the CNLSY, we pooled PPVT tests scores of 5- and 6-year-olds to get the first post–Head Start score for each child in our

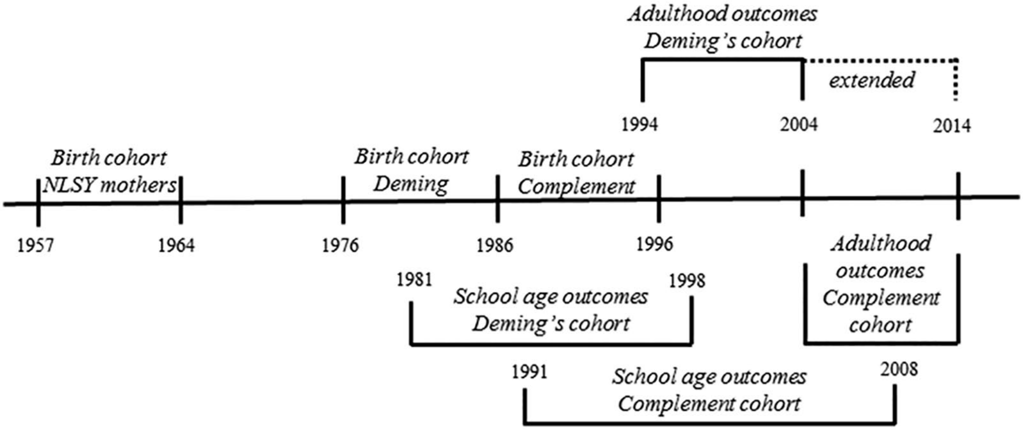

FIGURE 1. *Birth and outcomes time range by cohorts.*
Note. Time-wise, the combined cohorts sample (not shown) encompasses Deming's and complement cohorts. While boundary end points are approximate, they nonetheless include the bulk of each distribution: around 95% and 85% for Deming's cohort and complement cohort, respectively.

sample. Both the PIATMT and PIATRR were administered annually for respondents ages 5 to 14 and resulted in considerably more observations compared with the PPVT.

The two nontest outcomes covered ages 7 to 14 and included grade retention and learning disability diagnosis. As in Deming (2009), grade retention is a dichotomous variable based on survey respondents' answers to whether their child has ever been retained at grade level while in school. This question was asked biannually in the NLSY from 1988 to 2014. Grade retention was coded as a 1 if parents ever answered "yes" to this question across any of the survey years. Learning disability was based off a "yes" or "no" NLSY survey question that asked parents if their child had a learning disability. We coded our learning disability variable as 1 if respondents ever answered "yes" to this question, discounting a small number of children who were diagnosed with a learning disability prior to age 5.

Finally, as in Deming (2009), we included the same six young adulthood outcome variables: high school graduation, teen parenthood, some college attended, idleness, involvement with the justice system, and poor health status. All outcomes were measured up to the CNLSY 2014 survey round. Individuals were considered "idle" if they were not enrolled in school or had reported zero annual earnings—by 2004 for the Deming cohort and by 2014 for the complement cohort. The "involvement with the justice system" variable was constructed as a dichotomous variable, coded as one if a respondent ever answered "yes" to any survey question related to conviction, probation, sentencing, and prison. Teen parenthood was operationalized also with an indicator equal to one if a respondent's age at the birth of their first child was before 20 years old and applied to both female and male respondents. Finally, our poor health status variable was constructed by averaging a respondent's self-reported health framed by a 1-to-5 Likert-type scale (lower responses equating to poorer self-reported health). Poor health status was flagged by a dichotomous variable coded as one, if a respondent's average self-reported health score was less than three on the Likert-type scale.

Just as in Deming (2009), to reduce the risk of multiple-inference inflated Type I errors and mitigate measurement error, we constructed summary indices for the three test scores, an index for the two nontest score outcomes, and a final adulthood summary index (ASI) for the six young adulthood outcomes. Outcomes were normalized to have mean of zero and standard deviation of one, with positive index values signaling "good" outcomes and negative index values signaling "bad" outcomes. The final index was then created by taking a simple average of all the normalized and, where appropriate, resigned outcomes.

TABLE 1

Household Characteristics Averaged Over Head Start and No Preschool Status

	Head Start			No preschool				
Household characteristic	Deming's cohort	Complement cohort	Combined cohorts	Deming's cohort	Complement cohort	Combined cohorts	Diff. HS-none	Diff. HS-all
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)	−0.37/−0.49/−0.44	−0.50/−0.47/−0.46
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)	−0.26/−0.45/−0.41	−0.35/−0.53/−0.48
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)	0.01/0.01/ 0.02	0.15/0.14/0.18
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)	0.12/0.06/0.11	0.23/0.15/0.21
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)	0.08/−0.09/−0.02	−0.07/−0.28/0.21
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)	−0.04/−0.03/−0.06	−0.13/−0.14/−0.15
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)	−0.33/−0.63/−0.50	−0.47/−0.78/−0.66
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)	−0.31/−0.52/−0.43	−0.40/−0.58/−0.50
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/−0.22/−0.16	−0.23/−0.37/−0.33
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/−0.16/−0.18	−0.23/−0.22/−0.24
Sample size	779	637	1,491	1,931	1,857	3,658		
Sample size FE	435	475	972	769	1,098	1,799		

Note. Means and standard deviations were obtained for the full and sibling fixed effects samples, across cohorts. Differences in means (in *SD* units) between *Head Start* vs. *No preschool* status (Difference HS-None) and between *Head Star*t vs. *No preschool* + *Other preschool* status (Difference HS-All) were reported for Deming's cohort/Complement cohort/Combined cohort (in that order). Permanent income is the average over reported years of household net income (in 2014 dollars). The AFQT was age normed based on the NLSY79 empirical age distribution of scores, and then standardized ($M = 0$; $SD = 1$). Household characteristics for the *Other preschool* status group were documented in the Supplementary Table S2 in the online version of the journal. Mean differences between Deming's cohort and Complement cohort were all significant at the 1% level (with the exception of Grandmother's education; see Supplementary Table S2 in the online version of the journal). HS = Head Start; FE = fixed effect; AFQT = Armed Forced Qualification Test.

Comparing the distribution of these outcomes across the Deming and complement cohorts, we found evidence of substantial distributional shifts often favoring the complement cohort (see Supplementary Table S5 & Figure S6 in the online version of the journal). These between-cohort distributional changes might explain why Head Start impacts were dramatically different for the more recent set of siblings in the complement cohort. To illustrate, on the ASI, complement cohort siblings attending Head Start and not attending preschool were respectively 0.22 *SD* and 0.50 *SD* higher than the Deming cohort siblings. Similarly, although there was a small and marginally significant 0.10 *SD* difference on the cognitive test index across cohorts for Head Start attendees, that difference was more pronounced for the No Preschool status children (0.30 *SD; p* < .001). As these distributional shifts in outcomes were between cohorts across time and not within a cohort across Head Start treatment statuses, these results are not inconsistent with our later findings that the Deming and complement cohorts showed no signs of within-family selection into Head Start or No Preschool status (see Supplementary Table S7 in the online version of the journal).

Finally, 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, and on earnings.[12] The earnings composite for each sample member was obtained by first pooling all person-year earnings observations (in 2014 US$) and then regressing them on dummy-variable indicators for birth cohort and calendar year to purge earnings of birth cohort and measurement year effects.[13] 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 residuals for each individual, added them to the grand mean earnings in the sample, and took the natural logarithm of this earnings average.

Empirical Strategy

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. An FFE 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:

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

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 $\beta_1(\beta_2)$ 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. Next, $\mathbf{X}_{ij}$ represents the vector of "pretreatment" 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.

Selection Bias. Within-family comparisons remove the effects of time-invariant family characteristics on siblings' outcomes. There remains, however, a strong possibility of within-family selection bias. Reasons for different within-family patterns of care in early childhood 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 $\mathbf{X}$ in Equation 1—measured before or at the age of Head Start eligibility (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.[14]

Within our FFE framework, each covariate was thus regressed on siblings' preschool status, either Head Start or other 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 pretreatment characteristic and the regressed-on preschool status. These estimates are reported in the Supplementary Table S7 in the online version of the journal.

Focusing first on the complement cohort, siblings attending Head Start were on average older

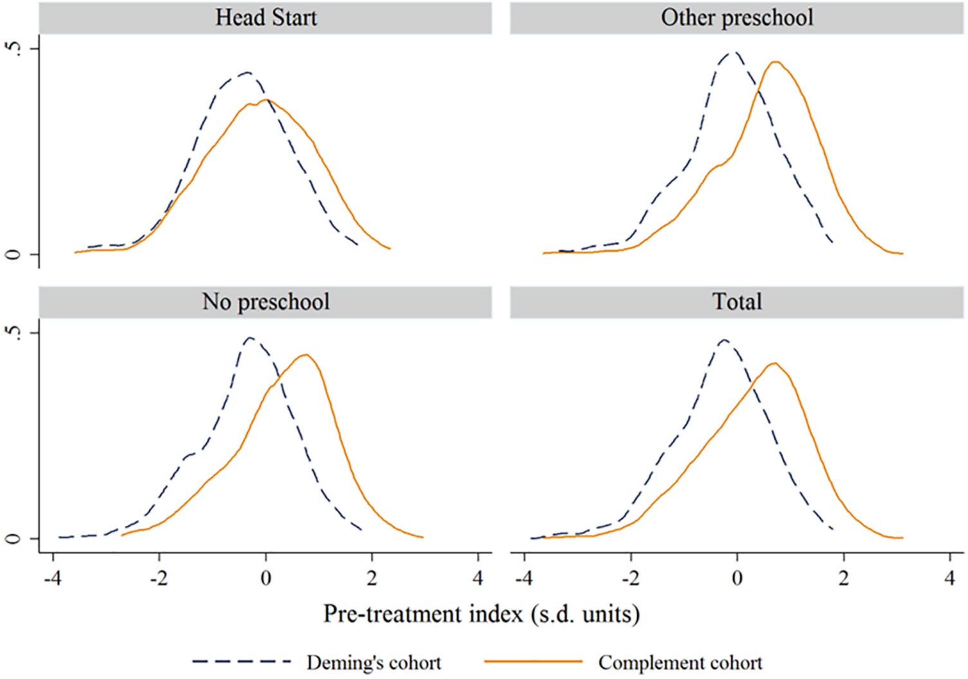

FIGURE 2. *Pretreatment index kernel density estimation by preschool status across cohorts.*
Note. Distributions were smoothed, and densities estimated via the Epanechnikov kernel function. Kolmogorov–Smirnov tests all indicated nonequality, at the 0.1% level, between compared densities.

by 1 year, and by almost 2 years for siblings attending other preschools, than their counterparts not attending any preschool program. This was consistent with the probability being greater for first-born sibling to be enrolled in preschool, by 10 *pp*s for Head Start enrollees, and by 28 *pp*s for other preschool participants. Both groups were 6 to 8 *pp*s less likely to receive maternal care from birth to age 3, and so more likely to receive care from a nonrelative (5 and 6 *pp*s, respectively). Attrition was low for both the complement and the combined cohorts, averaging about 4% and 3%, respectively. Moreover, both Head Start and other preschool participants were somewhat less likely (by about 3 and 2 *pp*s, respectively) to be part of observations lost to attrition.

To characterize selection bias as it pertains to overall disadvantage, a summary index of all pretreatment covariates was constructed in the same way as the multiple outcomes adulthood index described earlier.[15] For all cohorts, as with the covariates, the pretreatment index was regressed on the 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 pretreatment index were close to zero and never statistically significant. As with Deming's (2009) selection bias analysis (which we replicated, see Supplementary Tables S8–S9 in the online version of the journal), we could not reject the null hypothesis of equality between preschool statuses.

While 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 pretreatment index scores were shifted to the right, that is, toward 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 substitute for whatever impact the program would have otherwise yielded.

TABLE 2

Head Start Impacts on Cohorts' Adulthood Summary Index

Source of estimates	Model	Head Start	Other preschool	P value (HS = other)	R^2
Deming (2009)[a]	(1)	**0.14** (0.07)	0.08 (0.08)	.47	.12
Measurement period[b]: 1994–2004	(2)	**0.27** (0.08)	0.11 (0.08)	.12	.59
Sample size = 1,251 [364/364]	(3)	**0.23** (0.07)	0.07 (0.07)	.08	.62
Deming's cohort[c]	(1)	**0.14** (0.06)	0.08 (0.06)	.40	.21
Measurement period: 1994–2014	(2)	**0.18** (0.07)	0.02 (0.09)	.07	.64
Sample size = 1,251 [364/364]	(3)	**0.17** (0.07)	0.03 (0.07)	.13	.69
Complement cohort[d]	(1)	**−0.12** (0.06)	0.03 (0.06)	.01	.24
Measurement period: 2004–2014	(2)	−0.16 (0.10)	−0.05 (0.08)	.30	.71
Sample size = 2,144 [497/795]	(3)	**−0.15** (0.07)	−0.04 (0.05)	.15	.73
P value for model (3)[e] (Deming's = complement)		.01	.64		
Combined cohorts[f]	(1)	−0.02 (0.04)	0.06 (0.03)	.05	.24
Measurement period: 1994–2014	(2)	−0.01 (0.01)	0.01 (0.04)	.78	.61
Sample size = 3,738 [951/1,275]	(3)	−0.01 (0.04)	−0.003 (0.04)	.86	.63

Notes: Adulthood summary index (standardized) is a composite of six indicators: high school graduation, college attendance, teenage 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 pretreatment 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 pretreatment covariates. Model (3): same as model (2) with pretreatment covariates included. Standard errors are in parenthesis and clustered at the family level. Estimates in bold case were significant at the 5% level or less. HS = Head Start; AFQT = Armed Forced Qualification Test. [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 integrate both Deming's and the complement cohorts.

Results

Each of the following subsections is organized by cohorts. We first present Head Start impacts on the adulthood summary index (ASI), along with its individual composing outcomes. Second, longer-run Head Start impacts on ASI, educational attainment, college graduation, and earnings are described. Third, estimates for school-age outcomes are shown. Robustness checks and a reconciliation of results are presented in the final subsection.

Head Start Impacts on ASI

In Table 2, the FFE model was implemented in steps—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.[16] Model (1) included no FFEs but included the pretreatment covariates (see Supplementary Table S7 in the online version of the journal), 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 FFEs. Model (3) includes both FFEs and pretreatment 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, moving from model (2) to model (3) added some precision to the estimates, with the R^2 increasing from .64 to .69, from .71 to .73, and from .61 to .63 for the Deming, complement, and combined cohorts, respectively.[17]

From the middle panel, for the complement cohort, Head Start impacts on ASI were negative at −0.15 *SD* (significant at the 5% level). This value was in clear contrast ($p < .01$) with Deming's cohort estimate of 0.17 *SD* (*SE* = 0.07). In the bottom panel, for the combined cohorts model (3), no Head Start impact estimates were statistically significant; most were negative and close to zero.[18]

We investigated the decrease in Head Start impact on ASI from 0.23 *SD*—in Deming (2009); outcomes measured up to CNLSY 2004 survey round—to 0.17 *SD*; outcomes measured up to 2014. This change was due in part to the impact estimate on "Idle" changing sign and ceasing to be statistically significant. For example, while Head Start participants in Deming's cohort were 7 *pp*s (*SE* = 0.04) less likely to be "idle" in 2004, by 2014, the estimated impact disappeared (−3 *pp*s; *SE* = 0.04). Overall, 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 ASI.[19] For Deming's cohort, "Poor health status" stayed favorable by 5 *pp*s (*SE* = 0.03), decreasing slightly from 2004 (7 *pp*s; *SE* = 0.03). Impacts on "Some college attended" rose to statistical significance (11 *pp*s; *SE* = 0.04).[20] 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. As impacts on ASI are based on scores averaged across the composing indicators, Deming's cohort overall decline on this index were captured by changes on the individual outcomes just described.

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 (−0.08; *SE* = 0.03), and significant at the 1% level. Thus, in the complement cohort, siblings who have attended Head Start were less likely by about 8 *pp*s 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 (−0.07 *SD; SE* = 0.04), as well as for "Crime" (reversed scaled; −0.03; *SE* = 0.03).[21] In sum, the discrepancies between the two cohorts over Head Start impacts on these individual outcomes are aligned with the difference observed earlier over ASI (Table 2). Once more, impact estimates for the combined cohorts sample were small and never statistically significant.

Finally, consistent with M. L. 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 Supplementary Tables S10–S13 in the online version of the journal). Over the extended ASI (Deming's cohort), Head Start impact was estimated at 0.23 *SD* (*SE* = 0.11) for females versus 0.10 *SD* (*SE* = 0.10) for males.[22] Furthermore, females possibly carried more of Head Start's impact on educational attainment with an estimate at 0.34 *SD* (*SE* = 0.21) against 0.27 *SD* (*SE* = 0.21) for males (see Supplementary Table S13 in the online version of the journal).

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 Supplementary Table S13 in the online version of the journal). As described previously, impacts on ASI declined from Deming's (2009) published results 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, did not translate into either higher college graduation rate or to significantly higher adulthood earnings.[23]

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 & Pinto, 2015) was not the focus of this article, school-age outcomes

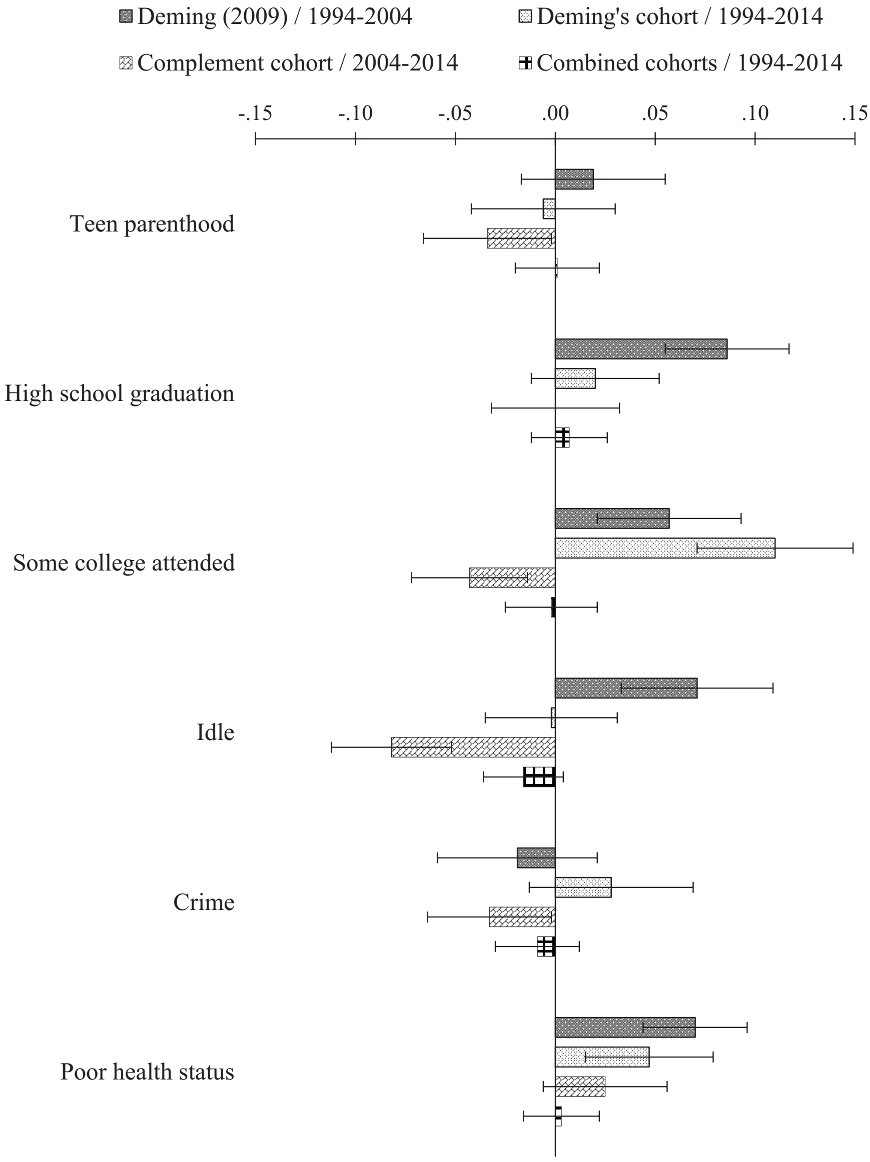

FIGURE 3. *Head Start impacts on the adulthood index individual outcomes across cohorts.*
Note. Measurement periods are displayed to the right of each label. Impacts are expressed as proportions. Deming's cohort, $N = 1,251$; complement cohort, $N = 2,144$; combined cohorts, $N = 3,768$. The counterfactual 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.

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.[24]

As shown in Figure 5, estimates from Deming (2009) and our replication of Deming (2009) 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 nontest index (−0.15 *SD; SE* = 0.08), with Head Start's impact on the learning disability diagnosis indicator (reverse scaled; −0.04 *SD*) being statistically significant at the 5% level. This is in line

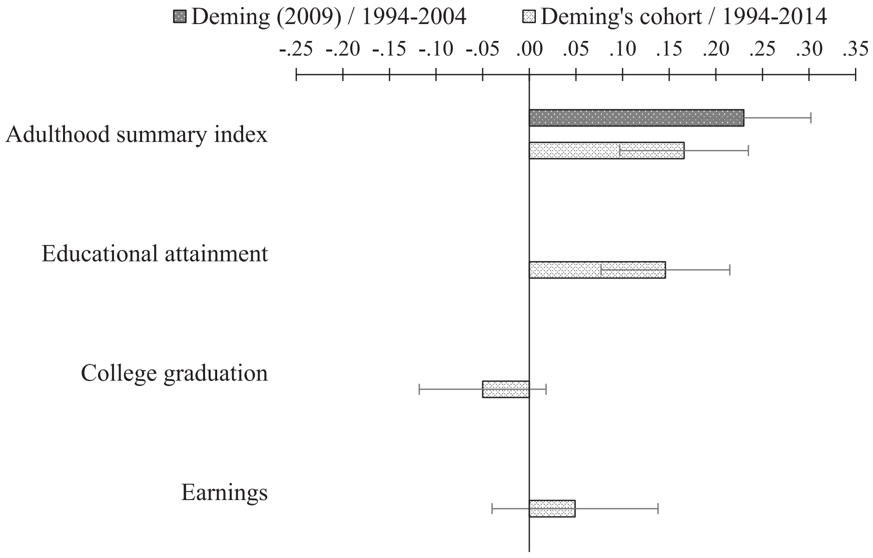

FIGURE 4. *Head Start longer-run impacts.*
Note. Measurement periods are displayed to the right of each label. Impacts are expressed in standard deviation units. For both cohorts, N = 1,251. Recall that the Deming (2009) study did not estimate Head Start impacts on educational attainment, college graduation, 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.

with the adverse impacts recorded on the Behavior Problems Index (reverse scaled; −0.07 *SD; SE* = 0.05). Yet, we could not detect any Head Start impact on "Grade retention" (whereas for Deming's cohort, the impact of being grade retained was at a lesser 7 *pp*s, significant at the 10% level).[25] Regarding the cognitive tests index, the relatively sustained gains generated by Head Start for Deming's cohort (0.11 *SD; SE* = 0.06) did not reflect those for the complement cohort (−0.02; *SE* = 0.06), while equality between the two estimates could not be rejected (p = .24). 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 to 6, 7 to 10, and 11 to 14 (see Supplementary Tables S14–S15 in the online version of the journal). Head Start impacts on the BPI index were stable across age groups and cohorts and were of similar magnitude as the 5 to 14 average.

Deming (2009) reported some fadeout of Head Start impact on the cognitive tests index by ages 11 to 14: from an estimate of 0.15 *SD* (*SE* = 0.09) by age period 5 to 6 to one of 0.06 *SD* (*SE* = 0.06). In contrast, for the complement cohort, a fadeout from a small but positive estimate (0.06 *SD; SE* = 0.07) might have occurred faster by age period 7 to 10 (−0.03 *SD; SE* = 0.06): The difference in impact with Deming's cohort for this age group (0.13 *SD, SE* = 0.06) was of marginal significance (p = .12). Complement cohort estimates ended at −0.05 *SD* (*SE* = 0.07) by age period 11 to 14. Finally, the combined cohorts sample faced a similar trend as the complement cohort; overall, impacts approached zero earlier for later Head Start cohorts.

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, we find that for both the Head Start and other preschool measures, impacts on cognitive outcomes were positive at treatment outset (age 5–6), followed by fadeout, possibly occurring earlier for Head Start participants (see model (5) in Supplementary Table S14 in the online version of the journal). Second, impacts on the nontest score index were unfavorable and statistically significant, for both

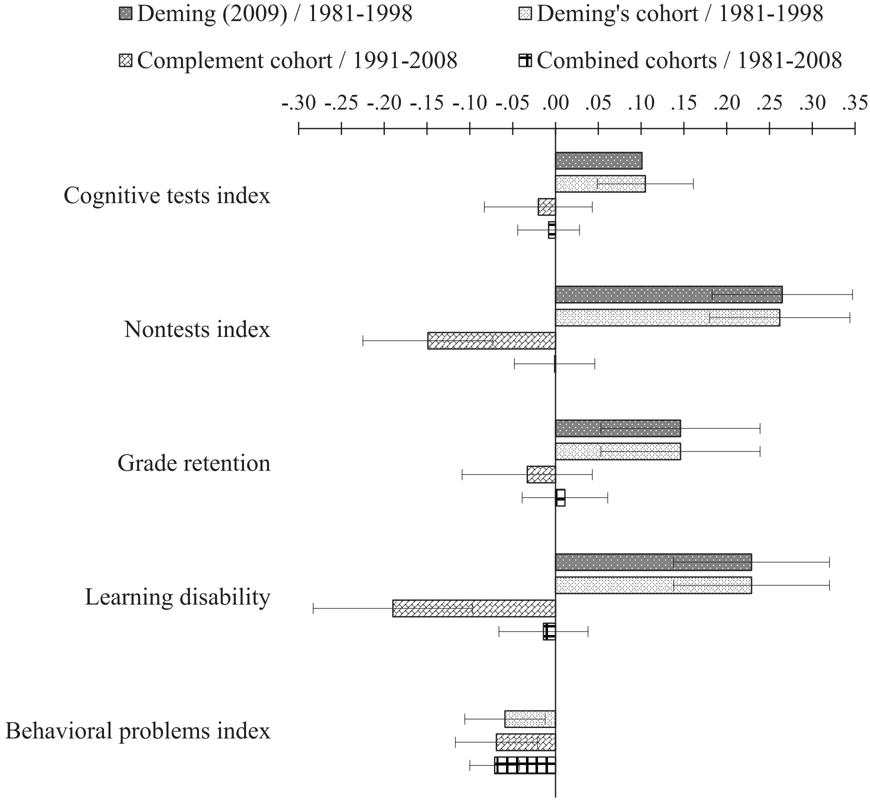

FIGURE 5. *Head Start impacts on school-age outcomes across cohorts.*
Note. Measurement periods are displayed to the right of each label. Impacts are expressed in standard deviation units. Deming's cohort, $N = 1{,}251$; complement cohort, $N = 2{,}144$; combined cohorts, $N = 3{,}768$. The counterfactual was 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 Behavior Problems Index (not statistically significant) but did not report them. "Grade retention" and "Learning disability" composed the "Nontest index."

preschool statuses (see Supplementary Table S10 in the online version of the journal). Third, impacts on the Behavior Problems Index were sometimes significant, mostly similar in magnitude, and unfavorable over all age periods across both preschool groups (see model (5) in Supplementary Table S15 in the online version of the journal). 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 (see respective top panels of Supplementary Tables S10–S13 in the online version of the journal).

Robustness Checks

As noted above, while the FFE design has been a workhorse empirical strategy to estimate the causal impact of Head Start (Bauer & Schanzenbach, 2016; Currie & Thomas, 1995; Deming, 2009; Garces et al., 2002), recent research has called into question this approach both in terms of the external (D. L. Miller et al., 2019) and construct validity of the FFE design (Heckman & Karapakula, 2019). The following section will discuss these threats and how, if at all, our primary findings change as a result.

Selection Into Identification. D. L. Miller et al. (2019) showed FEs can induce nonrandom selection of individuals into the FE identifying sample, leading to biased FE estimates relative to the average treatment effect (ATE) unless reweighting on observables is completed. In the FFE context, it may be that families with differential sibling participation in Head Start are systematically different across a variety of measures compared with those families with siblings that do not have variation in preschool status. If present, this "selection into identification" (SI) is a threat to the external validity of our results and could lead to a biased FFE estimate of Head Start's impact compared with the ATE (D. L. Miller et al., 2019).

To address this potential problem, following our above FFE analysis, we performed the reweighting-on-observables procedure discussed in D. L. Miller et al. (2019), first checking whether the combined cohorts FFE identifying sample exhibited SI across a variety of observables, including child's birth cohort, family size, mother's age at child's birth, permanent income, maternal AFQT, and whether a child was African American. Using the "switcher" and "non-switcher" naming convention found in D. L. Miller et al. (2019)—where (non-)"switcher" represents those families with (no) sibling variation in preschool status—and estimating propensity scores via multinomial logistic regression, we found differences over all family characteristics between Head Start participating "switcher" families and non–Head Start "non-switcher" families. By contrast, the only statistically significant difference ($p < .001$) in predictors between Head Start "switcher" versus "non-switcher" families (i.e., in which all siblings attended the program) was the indicator whether a child was African American.

Given that the FFE identifying sample for the combined cohorts exhibited some degree of SI, we corrected the potentially biased FFE estimates using the one-step reweighting-on-observables procedure in D. L. Miller et al. (2019). Overall, we found no evidence for the combined cohorts that reweighting changed the estimates of Head Start's impact on young adult outcomes, including high school graduation and ASI. We also performed the D. L. Miller et al. (2019) reweighting-on-observables procedure for the Deming cohort sample. Similar to D. L. Miller et al. (2019), we found reweighting attenuated the FFE estimates of Head Start for high school graduation, idleness, learning disability, and poor health, but found little change on other young adult outcomes or ASI. These results closely replicate D. L. Miller et al.'s (2019) findings and can be found in Supplementary Table S16 in the online version of the journal.

Spillovers. The construct validity of FFE estimates has also recently come into question. Revisiting the Perry Preschool Project, Heckman and Karapakula (2019) showed siblings who participated in the program had large positive spillovers on their nonparticipating siblings, particularly for male siblings. Although this problem is material to the FFE design, it is of less concern for this article as we were interested in comparing FFE results across cohorts and between short-run and long-run outcomes. In addition, it is unclear how sibling spillovers could explain this article's main findings, unless sibling spillover effects had become much stronger in the later as opposed to earlier cohorts.

That said, given Head Start's provision of wraparound services, some of which may influence parenting practices, some degree of spillover across children is likely (Deming, 2009; Ludwig & Miller, 2007; U.S. Department of Health and Human Services, Administration for Children and Families, 2019). Garces et al. (2002) and Deming (2009) test for Head Start treatment spillovers by interacting an indicator for first-born status with an indicator for Head Start treatment status. If spillover effects are present from an older to younger sibling, one would expect the impact of Head Start to be larger for non-first-born siblings (Deming, 2009). Consistent with Deming (2009), we found sparse and inconsistent evidence of spillovers. This holds true for both the Deming cohort and combined cohorts samples (see Supplementary Tables S17–S18 in the online version of the journal).

Reconciliation of Cross-Cohort Results

Human Capital Index. 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 (Cronbach's $\alpha = .83$) 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 for Head Start and other preschool from Equation 1. The interaction impact estimate on ASI, for the combined cohorts sample, was not statistically significant for any value within the range of the household human capital factor.

Cohort Covariates. To check whether the difference between Deming's cohorts and complement cohorts was due to some of the covariates we presented above, we interacted Head Start with an indicator for whether siblings belong to the complement cohort, along with interactions of Head Start with a series of covariates: first interacting one covariate at a time with the main effect included, then interacting all covariates with all main effects included. These covariates included pretreatment 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 1 *SD* below the mean. In addition, the 2007–2009 Great Recession could have negatively impacted complement cohort siblings who lived through its aftermath as teenagers. Thus, an indicator created for complement cohort siblings between age 12 and 18 in 2008 was also added. Had any of these interactions substantially reduced the estimate from Head Start × Cohort interaction, then these covariates would have explained some of the cross-cohort change in Head Start impacts between the Deming and complement cohort. We did not find any evidence that this was the case (see Supplementary Table S19 in the online version of the journal).

Blinder–Oaxaca Decomposition. A Blinder–Oaxaca decomposition approach (Blinder, 1973; Oaxaca, 1973) allowed us to consider how much a group mean difference on an outcome of interest (*Y*) is explained by group differences in predictors. Using the Deming cohort group (*d*) and complement cohort group (*c*), we formalized a threefold decomposition (Jann, 2008) as follows:

$$Yd - Yc = E + C + I$$

where $E = (Ed - Ec)$ measures the part of complement cohort's expected change in ASI, with complement cohort's predictors' means (i.e., endowments) fixed at Deming's cohort levels; $C = (Cd - Cc)$ measures the part of complement cohort's expected change in ASI, with complement cohort coefficients fixed at Deming's cohort levels; and $I = (Ed - Ec) \times (Cd - Cc)$ measures the contribution of the interaction between endowments' and coefficients' respective cross-cohort differences. The results from this analysis highlight key differences between Deming's and complement cohort samples, and showed how these differences—most notably in the pretreatment index and mother's age at child's birth—drive variation in estimated Head Start impacts across cohorts.

Breaking down the threefold decomposition results, we first found that the mean difference in ASI for Head Start attendees between the Deming cohort and complement cohort ($Yd - Yc$) was −0.22 *SD* (*SE* = 0.09).[26] The direction of this mean difference favored the complement cohort and was statistically significant. Second, having chosen the pretreatment index and mother's age at child's birth as predictors in the analysis,[27] we found that predictors' endowment parts of the decomposition (*E*) explained all of the outcome group mean difference (−0.27 *SD; SE* = 0.10), with mother's age at child's birth endowment recovering the near totality of it (−0.21 *SD; SE* = 0.10). The coefficient (*C*) and interaction (*I*) parts of the decomposition were negligible and not statistically significant (p >.84; p >.95). In sum, with mother's age at child's birth fixed at Deming's cohort level, complement cohort Head Start attendees would share an ASI expected value similar to that of their Deming's cohort counterparts.

Furthermore, if, for siblings of the complement cohort's counterfactual no preschool group, the expected change in ASI mean was also explained by mother's age at child's birth (as endowments, coefficients, or interaction effect), then, keeping this factor equal, Head Start would have had an impact of similar magnitude across cohorts. Thus, a threefold decomposition was conducted for counterfactual no preschool group siblings. Cross-cohort mean difference on the outcome of interest ($Yd - Yc$) was moderate and statistically significant (−0.56 *SD; SE* = 0.09). However, the *E* and *C* components of the decomposition were negligible (p >.88; p > .74), whereas the *I* component—that is, the interaction of mother's age at child's birth endowments' and coefficients' differences—recovered the outcome differential (−0.61 *SD; SE* = 0.18).

Because of the NLSY design, later cohorts of children had, on average, older mothers. Earlier cohorts (i.e., Deming's cohort) were born to relatively younger mothers, many of which may have disproportionately benefited from Head Start due to the program's provision of wraparound services that emphasize parental involvement (Currie & Neidell, 2007; Deming, 2009). This factor alone might have contributed to the discrepancy in our estimated Head Start impacts between Deming's and later cohorts. In particular, within Head Start participating families, siblings who did not attend a preschool program appeared to have on average benefited the most from having an older mother during their early years. Finally, the explanatory power of the mother's age at child's birth predictor remained robust after adding other covariates to the Blinder–Oaxaca decomposition, including permanent family income, whether the mother attended college (Table 2), a family human capital index, and family size (see *Notes* of Supplementary Table S19 in the online version of the journal).

Secular Trend. Our original cohort analysis looked only at the impact of Head Start across two cohorts—the Deming cohort and the complement cohort—obscuring whether over time there was a gradual secular decline or precipitous drop in the effect of Head Start. To address this, we decomposed our overall sample into three new birth-cohort-year groupings: C1—families with all siblings born before 1983, C2—1983–1987, and C3: post-1987. More than 90% of the C1 siblings and around a third of the C2 siblings were part of the Deming cohort sample. Two thirds of the C2 siblings and all of C3 siblings were part of the complement cohort sample.

Across virtually all of our outcomes, Head Start more favorably affected the C1 cohort compared with C2 and C3 cohorts. Moving from older to more recent cohorts, we observed unfavorable sign changes in the direction of Head Start's impact for a variety of short-term and longer-term outcomes, including cognitive and noncognitive school-age measures, ASI, educational attainment, and earnings.[28] In addition, the estimated trends indicate a comparatively larger decline in adulthood outcomes for individuals belonging to the most recent post-1987 cohort. In contrast, for school-year nontest outcomes, the decline was observed in much earlier cohorts. As reported in the Supplementary Table S20 in the online version of the journal, Head Start impacts between C1 and C3 on ASI went from positive and statistically nonsignificant (0.09 *SD; SE* = 0.09) to negative and significant at the 10% level (−0.25 *SD; SE* = 0.14). By contrast, Head Start impact on the nontest index went from a statistically significant 0.29 *SD* (*SE* = 0.10) advantage for C1 to a negligible −0.06 *SD* (*SE* = 0.15) for C3, with the drop occurring between C1 and C2 (−0.10 *SD; SE* = 0.15).

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 noncognitive and behavioral measures or on the adulthood summary index. Third, combining both cohorts produced Head Start impact estimates on all measured outcomes that were small and not statistically significant. Finally, we conducted a handful of empirical tests to determine what is driving the differences in Head Start impacts across birth cohorts. Although not definitive, our results suggest the important role factors outside of Head Start, notably mother's age at child's birth and the pretreatment index, played in moderating the program's impacts on participating children's young adult outcomes.

Given the benefit of time and 10 additional years of NLSY data, this article provides an updated estimate of Head Start's impact on observed adult wages for the Deming cohort. Deming (2009) originally estimated a 0.11 log points Head Start impact on adult wages; however, as survey participants were too young at the time of the calculation to report actual wages, this estimate was based on Deming's projection of future adult wages. Using the now available realized adult wages for siblings 25 years old or more, unconditional on employment and averaged across each survey round up to 2014, we follow up on his initial projection finding a nonstatistically significant smaller impact at 0.07 log points (*SE* = 0.12). Using dollar (2014 US$) instead of log earnings produces a negative and still nonstatistically significant impact of −US$999 (*SE* = US$1,507). In sum, Head Start generated no clear adult earnings gain for the Deming (2009) cohort siblings, although the large confidence interval on our estimate includes Deming's original estimated impact.[29]

Johnson and Jackson (2019) analyzed earlier cohorts of siblings born before 1976 with a dynamic complementarity design (i.e., capitalizing on two exogenous sources of variation separated in time). They found a larger, more precise estimate: Attending Head Start at age 4—that is, facing an average Head Start spending versus no spending, *coupled* with (and sensitive to) an average public K–12 spending—boosted earnings of poor children (measured from age 20–50) by 0.10 log points (*SE* = 0.02).[30] These positive estimates, generated from a different identification strategy and earlier cohorts, are close to Deming's projections, but the variation and imprecision associated with our estimates underscores the importance of monitoring cross-cohort trends in Head Start impacts.

Apart from revisiting and extending Deming (2009), this article adds to the current Head Start literature by finding heterogeneous program affects across CNLSY birth cohorts, perhaps due to cohort differences in resources that might serve as substitutes for Head Start. We find evidence of a notable rightward shift in the distribution of mothers' age at child's birth and in the pretreatment index across time, the latter indicating more favorable household conditions across time for mothers and their children in the years leading up to Head Start eligibility. As discussed in Deming (2009), this feature is an artifact of the CNLSY sampling design. Subsequent findings, primarily from our threefold Blinder–Oaxaca decomposition analysis, suggested that the changing effects of Head Start across cohorts were due largely to changes in covariates of the identifying sample, notably mother's age at child's birth, and not due to changing effectiveness of Head Start across time.

Although this study found that the counterfactual conditions of household characteristics (e.g., maternal age) predicted cross-cohort differences on long-term outcomes, additional factors may be driving these variations. Both the United States' substantial increase in spending on means-tested programs between the 1980s and 2010s and possible changes to program quality (e.g., due to a steep and continuous enrollment increase) could also explain these observed changes (Ludwig & Phillips, 2008), as could changes in labor market conditions and the return to specific forms of human capital over time. For example, 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 finding that Head Start negatively affected "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. Each of these explanations is potential avenue future research could take to further explore these results.

Overall, this article 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, necessitates further research on recent cohorts of Head Start attendees using complementary identification strategies.

Authors' Note

First authorship is shared between the first two authors, who made equal contributions to this publication.

Acknowledgments

We wish to thank Felipe Barrera-Osorio, Damon Clark, David Deming, Chloe Gibbs, Josh Goodman, and Eric Taylor for their feedback on prior drafts and presentations of this project.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. An overview of the literature on Head Start's short- and medium-term impacts is available in the Supplementary Section S1 in the online version of the journal.

2. For example, Johnson and Jackson (2019) calculated that for a child attending Head Start, a 10% increase in K–12 spending boosted educational attainment by 0.4 years, earnings by 20.6%, and reduced the probability of being incarcerated by 8 percentage points (*pp*s).

3. Nearly 95% of children included in the Deming (2009) family fixed effect (FFE) estimation sample were born between 1976 and1986. The remaining 5% of children were born between 1970 and 1975.

4. The adulthood summary index used in Deming (2009) included high school graduation, college attendance, teenage parenthood, idleness (i.e., neither working nor attending school), crime, and poor health status.

5. The FFE analysis sample used in this article includes 1970–1996 National Longitudinal Survey of Youth 1979 Children and Young Adults (CNLSY) birth cohorts.

6. Earnings are computed as the log of 1994–2014 averaged earnings, adjusted for age and survey year.

7. The Head Start impact estimates of Carneiro and Ginja (2014), Bauer and Schanzenbach (2016), and Barr and Gibbs (2019) were based on samples that included 1990 birth cohorts. However, these studies did not systematically estimate Head Start impacts by birth cohorts across time. Instead, primary results were 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.

8. 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 Rules 1 and 2); it is in that sense that this new cohort is the complement of Deming's cohort. Of all siblings comprising the complement cohort, 78% had reached 4 years of age post-1990 (75%, for Head Start participants). As in Deming (2009) and for all cohorts, the original National Longitudinal Survey of Youth 1979 (NLSY79) over sample of low-income Whites was excluded. One final point, the sum of Deming's cohort and the complement cohort is smaller than the size of the combined cohorts sample as there are more opportunities for siblings to meet the FFE eligibility criteria for the latter. For example, there were cases where families included in Deming's cohort had an additional child or children in later years that were not age eligible for Deming's cohort but were old enough to be a candidate for the complement cohort. However, to be included in the complement cohort, these age-eligible children needed to exhibit differential participation in Head Start. Thus, the family with one child was automatically disqualified from the complement cohort, and the family with multiple new children would also be excluded unless these children exhibited differential participation in Head Start. Regardless of their inclusion within the complement cohort, the children in each of these scenarios would be included in the combined cohorts sample.

9. 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.

10. Deming (2009) presented these characteristics over three preschool statuses (i.e., 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 also by racial/ethnic subgroups), see Supplementary Tables S2 through S4 in the online version of the journal. There was a reduction in sample sizes when restricting on families that had differential preschool status for at least two siblings (see "Data" subsection, Rule 2): by 66% for Deming's cohort, by 41% for the complement cohort, and by 45% 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).

11. Differences were even more pronounced when comparing between Head Start and the other preschool status (see Supplementary Table S2 in the online version of the journal).

12. 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.).

13. We opted to combine all respondent age 35+ into a single birth cohort dummy to ensure a sufficient sample size (N = 589) before regressing for adjustment. For all other birth cohorts, sample sizes were of at least 300 observations. The arguably arbitrary 35-year threshold was chosen such that *a priori* valid inputs would be available.

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

15. As in Deming (2009), variables comprising the pretreatment covariates index were all first positively oriented with respect to the adulthood summary index. For example, variables like gender (male), age (older), or grandmother 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, which in turn was also standardized ($M = 0$; $SD = 1$).

16. Deming (2009) used a similar approach in estimating impacts on the school-age cognitive tests index.

17. From similar trends, Deming (2009) concluded—based on Altonji et al.'s (2005) seminal work on the topic—that estimates obtained from model (3) stood as lower bounds for Head Start causal impacts.

18. Throughout (see Supplementary Tables S10–S13 in the online version of the journal), we considered identical demographic subgroups as in Deming (2009). Looking at the "Adulthood index" column in Supplementary Table S10 in the online version of the journal, 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—1 *SD* below NLSY79 AFQT empirical sample average (see bottom panel of Supplementary Table S10 in the online version of the journal). For these siblings ($N = 810$), Head Start appeared to have had a marginally significant positive impact on the adulthood summary index (0.11 *SD; SE* = 0.08). This estimate was greater for Deming's cohort (0.38 *SD*, significant at the 1% level). For the complement cohort, impact was much smaller (0.03 *SD, SE* = 0.15), although difference testing between this cohort's estimate and that of Deming's cohort did not fall below the 10% level of statistical significance ($p = .15$). 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).

19. For all estimates on individual outcomes, overall and by subgroups, see Supplementary Tables S11 and S12 in the online version of the journal.

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

21. 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's (2014) Head Start evaluation study. The estimate for the combined cohorts was very small though, with a standard error well balanced across zero (Figure 3; Supplementary Table S12 in the online version of the journal).

22. Although we could not reject equality between these estimates, we detected a favorable and statistically significant Head Start impact difference (see Supplementary Table S12 in the online version of the journal) of about 8 *pp*s 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-*pp* advantage on the "Crime" outcome (i.e., whether involved with the justice system).

23. Head Start impacts were positive and statistically significant for the subgroup of siblings whose maternal IQ (intelligence quotient) background was 1 *SD* below the *M*: 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 Supplementary Table S13 in the online version of the journal.

24. We conducted this section's analysis as in Deming (2009) and considered identical outcomes and age groups. The full set of estimates was compiled in the Supplementary Tables S12–S13 and S19–S20 in the online version of the journal.

25. Other preschool impact estimates had similar trends on all these "noncognitive" outcomes (see Supplementary Tables S12–S13 in the online version of the journal).

26. In this complementary analysis, standard errors were clustered at the family level.

27. These predictors were selected based on both the magnitude and direction of the mean covariate differences between the Deming and complement cohort. Mother's age at child's birth was around 7 years higher, and the mean pretreatment index statistically more favorable (0.20 *SD; SE* = 0.08) for complement cohort's Head Start attendees.

28. 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.

29. Earnings (log transformed or not) regression estimates were very similar—0.05 log points; *SE* = 0.11; −1030 2014 US$; *SE* = 1488, respectively—whether age was controlled for instead of year of birth. As described in the "Results" section, for Deming's cohort,

attending Head Start versus no preschool yielded 0.3 years increase of completed schooling. From this, and based on Card's (1999) review on returns to education of about 5% to 10% per year of completed schooling, we might have expected to find evidence of an impact on earnings between 1% and 3.5%.

30. 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 0.03 log point (*SE* = 0.03) and was no longer statistically significant at the 10% level. In contrast, with a 10% increase in K–12, the estimate jumped to 0.17 log point ($p < .01$).

References

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *The Journal of Political Economy*, *113*(1), 151–184.

Anderson, K., Foster, J., & Frisvold, D. (2010). Investing in health: The long-term impact of Head Start on smoking. *Economic Inquiry*, *48*(3), 587–602.

Anderson, M. 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.

Augustine, J. M., Prickett, K. C., Kendig, S. M., & Crosnoe, R. (2015). Maternal education and the link between birth timing and children's school readiness. *Social Science Quarterly*, *96*(4), 970–984.

Bailey, M. J., Sun, S., & Timpe, B. (2018). *Prep school for poor kids: The long-run impacts of head start on human capital and economic self-sufficiency* [Unpublished manuscript]. Department of Economics, University of Michigan.

Barr, A., & Gibbs, C. R. (2019). *Breaking the cycle? Intergenerational effects of an anti-poverty program in early childhood* [Manuscript submitted for publication]. Annenberg Institute, Brown University.

Bauer, L., & Schanzenbach, D. W. (2016). *The long-term impact of the Head Start program* [Policy report]. The Hamilton Project. http://www.hamiltonproject.org/papers/the_long_term_impacts_of_head_start?_ga=2.58041645.688673869.1554397866-1708819741.1554397866

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start* [NBER Working Paper Series No. 20434]. https://www.nber.org/papers/w20434.pdf

Blinder, A. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, *8*(4), 436–455.

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1801–1863). Elsevier.

Carneiro, P., & Ginja, R. (2014). Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start. *American Economic Journal: Economic Policy*, *6*(4), 135–173.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593–1660.

Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, 1, 697–812.

Currie, J., & Almond, D. (2011). Human capital development before age five. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 4, pp. 1315–1486). Elsevier.

Currie, J., & Neidell, M. (2007). Getting inside the "black box" of Head Start quality: What matters and what doesn't. *Economics of Education Review*, *26*(1), 83–99.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *The American Economic Review*, *85*(3), 341–364.

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

De Haan, M., & Leuven, E. (2020). Head Start and the distribution of long-term education and labor market outcomes. *Journal of Labor Economics*, *38*(3), 727–765.

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

Duncan, G. J., Lee, K. T. H., Rosales-Rueda, M., & Kalil, A. (2018). Maternal age and child development. *Demography*, *55*(6), 2229–2255.

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

Frisvold, D. E. (2006). *Head Start participation and childhood obesity*. [Vanderbilt University Economics Working Paper No. 06-WG01]. https://ssrn.com/abstract=887433.

Frisvold, D. E., & Lumeng, J. C. (2011). Expanding exposure: Can increasing the daily duration of head start reduce childhood obesity? *Journal of Human Resources*, *46*(2), 373–402.

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review*, *92*(4), 999–1012.

Heckman, J. J., & Karapakula, G. (2019). *Intergenerational and intragenerational externalities of the Perry Preschool Project* [NBER Working Paper Series No. 25889]. https://www.nber.org/papers/w25889.pdf

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

Heckman, J. J., & Pinto, R. (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.

Hotz, V. J., McElroy, S. W., & Sanders, S. (2005). Teenage childbearing and its life cycle consequences: Exploiting a natural experiment. *Journal of Human Resources*, *40*(3), 683–715.

Hoynes, H., Schanzenbach, D. W., & Almond, D. (2016). Long-run impacts of childhood access to the safety net. *American Economic Review*, *106*(4), 903–934.

Jann, B. (2008). The Blinder–Oaxaca decomposition for linear regression models. *The Stata Journal: Promoting Communications on Statistics and Stata*, *8*(4), 453–479.

Johnson, R., & Jackson, C. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy*, *11*(4), 310–349.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, *122*(1), 159–208.

Ludwig, J., & Phillips, D. A. (2008). Long-term effects of Head Start on low-income children. *Annals of the New York Academy of Sciences*, *1136*(1), 257–268.

Miller, A. R. (2011). The effects of motherhood timing on career path. *Journal of Population Economics*, *24*(3), 1071–1100.

Miller, D. L., Shenhav, N. A., & Grosz, M. Z. (2019). *Selection into identification in fixed effects models, with application to Head Start* [NBER Working Paper Series No. 26174]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3442715#

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, *14*(3), 693–709.

Thompson, O. (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*. https://hses.ohs.acf.hhs.gov/pir/

U.S. Department of Health and Human Services, Administration for Children and Families. 2019). *Head Start program facts: Fiscal year 2019*. https://eclkc.ohs.acf.hhs.gov/about-us/article/head-start-program-facts-fiscal-year-2019

Zigler, E., & Valentine, J. (1979). *Project Head Start: A legacy of the war on poverty*. The Free Press.

Authors

REMY PAGES is a PhD candidate in the educational policy & social context area in the School of Education at the University of California, Irvine. His research concentrates on the evaluation of educational interventions in promoting persistent effects, with a current focus on college experience success measurement.

DYLAN J. LUKES is a PhD candidate in education policy and program evaluation at Harvard Graduate School of Education and Harvard Graduate School of Arts & Sciences. He studies the economics of education with a focus on early childhood education and technology & learning.

DREW H. BAILEY is an associate professor in the School of Education at the University of California, Irvine. His research focuses on understanding the longitudinal stability of individual differences in children's mathematics achievement and on the medium- and long-term effects of educational interventions. His current work attempts to use psychological theories and methods to build models to improve the accuracy of predictions about the medium- and long-term effects of educational interventions.

GREG J. DUNCAN is a Distinguished Professor in the School of Education at the University of California, Irvine. He has published extensively on child poverty and the importance of early academic skills, cognitive and emotional self-regulation as well as health in promoting children's eventual success in school and the labor market.

Manuscript received April 6, 2019
First revision received January 27, 2020
Second revision received July 7, 2020
Accepted July 8, 2020