American Economic Journal: Applied Economics 2009, 1:3, 111–134
http://www.aeaweb.org/articles.php?doi=10.1257/app.1.3.111

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

By David Deming*

This paper provides new evidence on the long-term benefits of Head Start using the National Longitudinal Survey of Youth. I compare siblings who differ in their participation in the program, controlling for a variety of pre-treatment covariates. I estimate that Head Start participants gain 0.23 standard deviations on a summary index of young adult outcomes. This closes one-third of the gap between children with median and bottom quartile family income, and is about 80 percent as large as model programs such as Perry Preschool. The long-term impact for disadvantaged children is large despite "fade-out" of test score gains. (*JEL* H52, J13, I28, I38)

Head Start is a federally funded and nationwide preschool program for poor children. Started in 1965 as part of the "War on Poverty," it serves over 900,000 children today and has funding of $6.8 billion annually (Office of Head Start 2008). Public investment in Head Start has more than tripled in real terms ($2.1 billion to $6.8 billion in 2007 dollars) since its inception, and in the past decade there has been an expansion in state-funded preschool programs (W. Steven Barnett et al. 2007). Still, despite substantial growth in Head Start enrollment over time, as one recent survey notes, "skepticism about the value of the program persists." (Jens Ludwig and Deborah A. Phillips 2008.)

This paper provides evidence of the long-term benefits of Head Start for a more recent birth cohort of children, most of whom were enrolled in the program between 1984 and 1990. My data source is the National Longitudinal Mother-Child Supplement (CNLSY), which surveyed the mothers of the National Longitudinal Survey of Youth (NLSY) 1979 every two years from 1986 until 2004. Tracking a cohort of Head Start participants over time has several advantages. First, the survey contains extensive information on family background from multiple members of the same families, allowing for intergenerational and

* Kennedy School of Government, Harvard University, 79 JFK St., Cambridge, MA 02139 (e-mail: demingd@nber.org). I would like to thank Professors Greg Duncan, Susan Dynarski, Thomas Kane, Jeffrey Liebman, Jens Ludwig, and Lawrence Katz for reading drafts of this paper and for providing essential guidance and feedback. I benefited from the helpful comments of Richard Murnane, Tristan Zajonc, and seminar participants at the National Head Start Research Conference, the Association for Public Policy Analysis and Management Meetings, the Harvard Economics Labor Lunch, and the Work in Progress seminar at the Kennedy School of Government. I gratefully acknowledge funding from the American Education Research Association, the Julius B. Richmond Fellowship at the Center for the Developing Child, the Taubman Center for State and Local Government, and the Multidisciplinary Program on Inequality and Social Policy at Harvard.

† To comment on this article in the online discussion forum, or to view additional materials, visit the articles page at: http://www.aeaweb.org/articles.php?doi=10.1257/app.1.3.111.

within-family comparisons. Second, participation was reported contemporaneously, mitigating concerns about measurement error or recall bias (Eliana Garces, Duncan Thomas, and Janet Currie 2002, henceforth GTC). Third, low attrition and consistent administration of tests and survey questions over time facilitate an analysis of the effect of Head Start over the life cycle.

Like Currie and Thomas (1995), henceforth CT, and GTC (2002), I identify the effect of Head Start using within-family differences in program participation. I control for a variety of pre-treatment characteristics, such as maternal work history, child care arrangements, and birth weight, which vary among members of the same family. Although some concerns remain in the absence of a randomized experiment, I show that there is little evidence of systematic bias in assignment to Head Start within families. I also show that controlling for these pre-treatment covariates has very little impact on the point estimates.

I find that the long-term impact of Head Start is about 0.23 standard deviations on a summary index of young adult outcomes, with larger impacts for relatively disadvantaged children. This gain is equivalent to about one-third of the gap between the bottom permanent income quartile and the median in the CNLSY sample, and is about 80 percent as large as the gains from the Perry Preschool and Carolina Abecedarian "model" preschool programs (Pedro M. Carneiro and James J. Heckman 2003; Michael L. Anderson 2008). The results are robust to different constructions of the summary index, different age and sample restrictions, and alternative definitions of Head Start participation.

This paper also sheds some light on the life-cycle benefits of early skill formation. Although nearly all school-age interventions use test scores as a benchmark for success, the connection between test score gains and improvements in adult outcomes is not well understood. More practically, without some sense of the connection between short- and long-term benefits, researchers must wait at least 15–20 years to evaluate the effect of an early childhood program. I find an initial age (5–6 years old) test score gain of about 0.15 standard deviations that fades out to less than half that amount by ages 11–14. Fade-out is particularly strong for African American children and very disadvantaged children. Still, it is these children who experience the largest long-term benefits. For children whose mothers scored one standard deviation below the average on a cognitive test score, the long-term effect of Head Start is 0.28 standard deviations, and yet their net test score gain is essentially zero. Thus, a projection of future benefits for these children based solely on test score gains would greatly understate the impact of the program.

These results also suggest that large returns to investment in early education are possible without costly scaling up of model, highest-quality preschool programs. A rough comparison indicates that Head Start generates about 80 percent of the benefits of these programs at about 60 percent of the cost (Anderson 2008; Currie 2001). Furthermore, plugging the effect sizes here into calculations of the social cost of high school dropout performed by Henry M. Levin et al. (2007) suggests a similar benefit-cost ratio between Head Start and Perry Preschool. The largest potential difference is in the effect of each on crime, which generates two-thirds of the estimated social benefit in Perry Preschool (Clive R. Belfield et al. 2006). In contrast, I find no impact of Head Start on criminal activity.

The rest of the paper is organized as follows. Section I provides background on Head Start and the related literature. Section II describes the data. Section III outlines my empirical strategy, and discusses issues of multiple inference and selection bias. Section IV presents the results and some robustness checks. Section V discusses comparability and engages in a speculative benefit-cost calculation. Section VI concludes.

I. Background

Head Start's mission is to promote school readiness "by enhancing the social and cognitive development of children through the provision of educational, health, nutritional, social, and other services to enrolled children and families" (Office of Head Start 2008). In practice, the program focuses on "whole child" development rather than academic preparation for kindergarten through content instruction, although there have been calls in recent years to make the program more academically oriented (Ron Haskins 2004). Funding guidelines require that 90 percent of participants be at or below the federal poverty level, on public assistance, or be foster children (Office of Head Start 2008). Thus, high-quality preschool education is beyond the means of most participating families. In addition to preschool education, Head Start provides services such as medical, dental, and mental health care (including nutrition), and child development assistance and education for parents. Head Start is a 9 month full- or part-day program, and it currently costs between $7,000 and $9,000 dollars per child, per year, or about 60 percent of model preschool programs to which it is often compared (Currie 2001).[1] Enrollment in the program is limited to a maximum of two years per child, although, in practice, the modal time of enrollment is one year (United States Department of Health and Human Services (HHS) 2005).

Head Start programs are administered locally, but quality is regulated by a set of federal guidelines (Office of Head Start 2008). Still, there may be considerable heterogeneity in implementation and program features across geographic areas and over time. Since overall preschool enrollment has increased greatly over the past 40 years, there is a much wider variety of alternatives for parents today than for earlier cohorts. If these alternatives are of increasingly higher quality, then, holding program quality constant, the impact of Head Start would fall over time. Similarly, since the measured impact of the program is implicitly a comparison with the child's counterfactual environment, children from more disadvantaged backgrounds may experience larger gains from Head Start (HHS 2001; GTC 2002; William T. Gormley, Jr. and Ted Gayer 2005). On the other hand, some studies find evidence of fade-out for African American participants compared to their more advantaged white peers (CT 1995; HHS 2005). A possible explanation is differences in school quality by race or

[1] The lower bound estimate comes from Head Start program Fact Sheet FY 2007 (http://www.acf.hhs.gov/programs/ohs/about/fy2007.html)(accessed May 22, 2009). The upper bound includes the estimated effect of state and local matching grants in addition to the federal funds reported on the HHS Web site. I thank Jens Ludwig for providing me with this data.

socioeconomic status (CT 2000). This implies that without follow-up, gains from an intervention are quickly lost.

Given that fade-out is an empirical regularity in educational settings in the United States (CT 1995; Alan B. Krueger and Diane M. Whitmore 2001; Anderson 2008) and internationally (Michael Kremer, Edward Miguel, and Rebecca Thornton 2004; Abhijit V. Banerjee et al. 2007), the use of test scores as a proxy for long-term measures of success is questionable. More concretely, if fade-out generalizes to all long-term impacts, the benefits of many of these interventions have been overstated. However, studies of model preschool interventions find dramatic improvements in long-term outcomes among program participants, despite rapid fade-out of test score gains (Frances A. Campbell et al. 2002; Lawrence Schweinhart et al. 2005; Anderson 2008).

The best evidence for the long-term impact of Head Start comes from two recent studies. GTC (2002) use the Panel Study of Income Dynamics to compare siblings in the same family who differ in their participation in Head Start. Ludwig and Douglas L. Miller (2007) use a discontinuity in Head Start funding across counties to identify the impact of additional funding on child outcomes. Using different data sources and identification strategies, each finds long-term impacts of Head Start on outcomes such as educational attainment, crime, and mortality, with some heterogeneity in subgroup impacts.[2] Both studies evaluate the effect of Head Start on cohorts enrolled in the program between 1965 and 1980. In neither case were short-term test score measures available, either because information on Head Start participation was based on retrospective reporting or because the identification came from county-level aggregates.[3]

II. Data

The original NLSY began in 1979 with 12,686 youths between the ages of 14 and 22. In 1986, the National Longitudinal Survey (NLS) began a separate survey of the children of the 6,283 women in the NLSY. As of 2004, these women had given birth to 11,428 children, mostly in the early years of the survey, since by 2004 many were beyond childbearing age. The CNLSY tracks every child born to an NLSY respondent, enabling a comparison of siblings within the same family. Furthermore, mothers are surveyed extensively prior to the birth of their children, which allows for a rich set of controls for early life circumstances. In every survey year beginning in 1988, mothers were asked if their children had ever participated (or were currently enrolled) in Head Start, and if they were enrolled in any other preschool. The NLSY included an oversample of the poor, so the children in the survey are more disadvantaged than average (US Department of Labor 2008).

I first restrict the sample to children who were over four years old by 1990. This ensures that all children in the sample are properly categorized (that is, they will

[2] GTC find impacts on educational attainment for whites only, and on crime for blacks only. Ludwig and Miller find roughly equal impacts for whites and blacks on educational attainment and age five to nine mortality.

[3] Ludwig and Miller link their data to the National Educational Longitudinal Survey of 1988 (NELS 1988) and the 2000 follow up, but find no evidence of test score gains for children in the treated counties.

not subsequently enroll in Head Start because they are no longer eligible). By 2004, these children are age 19 or older. The original NLSY included an oversample of low-income whites, but it was dropped in 1990 for budgetary reasons. Although this sample was included in CT (1995), I exclude it here. Finally, I restrict the sample to families with at least two age-eligible children. Together, these restrictions produce a sample size of 3,698.

Table 1 presents selected characteristics of the sample, separately for white/Hispanic children and for African Americans. The first row of means for each variable is for the age-eligible sample discussed above. Head Start participants of all races come from relatively disadvantaged backgrounds. They have lower permanent incomes, lower maternal Armed Forced Qualification Test (AFQT) scores, and lower levels of education.[4] The degree of negative selection is much greater for the white/Hispanic sample, however. Permanent income is 0.39 standard deviations lower for whites and Hispanics, but only 0.11 standard deviations lower for African Americans, relative to the "no preschool" sample.[5] This pattern is similar for maternal AFQT score, and is even stronger when comparing Head Start participants to children in other preschools.

The second row for each variable in Table 1 presents results for children in families where siblings differentially participate in Head Start, other preschools, or no preschool. In this paper, the impact of Head Start is identified by comparing siblings in the same family who vary in their participation in preschool programs. Thus, if all three of a mother's children were enrolled in Head Start, I cannot say anything about the effectiveness of the program for them. A comparison of the first and second rows for each variable gives a sense of the representativeness of this subsample. Not surprisingly, family differences across the three options narrow slightly. Again, however, whites and Hispanics are much more negatively selected into Head Start than African Americans, who are roughly equal to those who did not attend preschool in terms of income, maternal education, and AFQT score.

The generalizability of this sample to the population eligibile for Head Start is an important question. Using administrative data matched to the CNLSY, Currie and Matthew Neidell (2007) examine the characteristics of Head Start programs in the counties where surveyed children reside and find that they are generally representative of nationwide programs, although they are slightly larger, more urban, and have a higher percentage of nonwhite children. One thing to note is that children who are old enough for an examination of long-term impacts were often born to the younger mothers in the NLSY cohort. Younger mothers may benefit disproportionately from Head Start, since the program makes an active effort to involve and educate parents (Currie and Neidell 2007).

Outcomes.—First, I examine test score outcomes for children ages five and six, immediately following their enrollment in Head Start. Ideally, I would look at age

[4] Permanent income is constructed as the average of net family income from 1979 to 2004 in constant 2004 dollars. Maternal AFQT is a standardized test administered to NLSY respondents in 1981. To account for differences in the age at which the test was taken, I scale the score by the empirical age distribution in the full NLSY sample and normalize it to have a mean of zero and a standard deviation of one.

[5] Standard deviations are calculated separately for the white/Hispanic and black samples.

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

	White / Hispanic			Black			Head start—none diff. (in SD units)	
	Head Start (1)	Preschool (2)	None (3)	Head Start (4)	Preschool (5)	None (6)	White/ Hispanic (7)	Black (8)
Permanent income	26,553 [19,555]	52,130 [34,577]	35,592 [23,460]	24,005 [16,103]	32,470 [21,939]	25,980 [18,496]	–0.39	–0.11
Fixed effects subsample	27,560 [22,902]	41,882 [22,403]	35,901 [23,600]	26,010 [19,559]	28,940 [22,853]	24,164 [16,314]	–0.35	0.11
Mother < high school	0.51 [0.50]	0.18 [0.38]	0.42 [0.49]	0.33 [0.47]	0.20 [0.40]	0.38 [0.49]	0.18	–0.10
Fixed effects subsample	0.53 [0.50]	0.25 [0.43]	0.41 [0.49]	0.39 [0.49]	0.27 [0.45]	0.37 [0.48]	0.24	0.04
Mother some college	0.22 [0.41]	0.41 [0.49]	0.23 [0.42]	0.31 [0.46]	0.50 [0.50]	0.28 [0.45]	–0.02	0.07
Fixed effects subsample	0.16 [0.37]	0.31 [0.46]	0.22 [0.41]	0.32 [0.47]	0.42 [0.50]	0.30 [0.46]	–0.15	0.04
Maternal AFQT	–0.44 [0.73]	0.23 [0.85]	–0.21 [0.86]	–0.75 [0.49]	–0.51 [0.72]	–0.68 [0.60]	–0.27	–0.12
Fixed effects subsample	–0.48 [0.70]	0.02 [0.83]	–0.20 [0.82]	–0.77 [0.48]	–0.63 [0.66]	–0.76 [0.56]	–0.34	–0.02
Grandmother's education	8.53 [3.50]	10.62 [2.92]	9.34 [3.36]	9.71 [2.56]	10.88 [2.68]	9.70 [2.87]	–0.24	0.00
Fixed effects subsample	8.51 [3.42]	10.09 [3.19]	9.54 [3.34]	9.82 [2.59]	10.13 [2.76]	9.98 [2.67]	–0.31	–0.06
Sample size	364	745	1,374	415	249	551		
Sample size — FE	229	315	510	206	144	259		

Notes: Means and standard deviations are presented separately for the full and fixed effects sample, which consists of families where at least one sibling (but not all) participated in Head Start or other preschools. Permanent income is obtained by averaging reported family income (scaled to 2004 dollars) over the years for which data were available. AFQT scores are age normed according to the empirical age distribution of scores in the full NLSY sample, and then standardized to have a mean of zero and a standard deviation of one.

five test scores only, but the biannual survey design of the CNLSY means that around half of the children take tests at odd-numbered ages and the other half at even-numbered ages. So pooling five and six year olds ensures that I obtain the first post-program score for every child in the sample. The three tests analyzed are the Peabody Picture Vocabulary Test (PPVT), the Peabody Individual Achievement Math (PIATMT) subtest, and Reading Recognition (PIATRR) subtest.[6] All three are widely used and validated tests of cognitive function and/or academic achievement of children.[7] The PIAT subtests were administered every survey year for those

[6] There is also a PIAT Reading Comprehension (PIATRC) subtest. Administration of the test was conditional on a minimum score on the PIATRR, however, so I do not use this test.

[7] The score is reported as a nationally normed percentile score (from 0–99) that is age-specific, and so increases in the test scores are relative to the national out-of-sample average for children of the same age. The tests are scored and administered using Item Response Theory, and they are designed for children of varying ages.

age 5–14. Thus, by 2004, children had taken the PIATMT and PIATRR as many as five times each.[8] CT found a significant and persistent increase on the PPVT for white and Hispanic children, but found test score gains for African American children faded out by about age ten (CT 1995). In results not reported here, I replicate their specification successfully.

I also report results for two additional school-age outcomes: grade retention and the diagnosis of a learning disability. In each survey year, parents were asked if their child had been retained in any grade in school. The question was asked every year from 1988 to 2004, and so I construct an indicator variable that is equal to one if the child's parents answered "yes" in any survey year.[9] The second outcome is the diagnosis of a learning disability. Unfortunately, more information on the specific nature of the child's learning disability was unavailable. However, there is a separate category of health condition called "hyperactivity/hyperkinesis," which may rule out attention deficit hyperactivity disorder (ADHD) and related behavioral problems.[10] Although some parents and educators think of learning disabilities as genetic (or at least determined prior to school entry), diagnosis has increased significantly in the past 30 years, too fast for genes to be the sole factor (G. Reid Lyon 1996). The existence of a learning disability for a given child may be partly predetermined, but it could be diagnosed more readily if a child is unprepared for school (or is in a lower-quality school) and struggles with basic skills.

I examine the impact of Head Start on six different young adult outcomes: high school graduation, college attendance, "idleness," crime, teen parenthood, and health status. They were chosen to represent different outcome domains based on a priori notions of importance and in concordance with other studies of the long-term impact of early childhood intervention (Campbell et al. 2002; Schweinhart et al. 2005; Anderson 2008). Since there is some evidence that the General Educational Development (GED) certification is not rewarded equivalently to a high school diploma in the labor market (Heckman and Yona Rubenstein 2001), I also look at non-GED high school graduation. Respondents are considered idle if they are not enrolled in school and report zero wages in 2004. Crime is a composite measure of self-reported contact with the criminal justice system.[11] Teen parenthood applies to respondents of either gender, and excludes the small number of respondents who indicate that they were married at the time of childbirth. Finally, I measure self-reported health status by averaging responses to a Likert scale item on self-reported health status. Self-reported health status is a powerful predictor of mortality and other negative health outcomes even when controlling for doctor reports and other

[8] Unlike the PIAT, the PPVT was generally administered once between the ages of three and five, and once again after age ten, although there was considerable variation in the age at administration over time (US Department of Labor 2008). As a result, there are many less observations for the PPVT, and the panel of test scores by age is not balanced.

[9] Unfortunately, the grade in which the child was retained is not available in every survey year. Since the survey was administered every two years, it is difficult to determine when retention occurred. Evidence from CT (1995), and confirmed in analyses not reported here, however, suggests that almost all of the effect of Head Start on grade retention occurred by age ten.

[10] There was no effect of Head Start on this outcome, nor was there an effect on the Behavior Problems Index (BPI), a commonly used measure of age-appropriate behavior problems.

[11] Specifically, it is an indicator variable equal to one if the respondent reports ever having been convicted of a crime, been on probation, sentenced by a judge, or is in prison at the time of the interview.

behaviors (Anne Case, Darren Lubotsky, and Christina Paxson 2002). I generate an indicator variable that is equal to one if the average response is less than three out of five, or "fair health."

III. Empirical Strategy

A. *Multiple Inference*

I assess the impact of Head Start on three different test scores that are administered over multiple years, and two school-age and six young adult outcomes. Furthermore, because past research has found large and important differences in the effect of early childhood intervention by race (CT 1995), gender (Anderson 2008), and socioeconomic status (HHS 2001; Gormley and Gayer 2005), I would ideally look at the effect of Head Start for these groups separately. With this many outcomes and subgroups, and a relatively small sample size, multiple inference can be an important issue. To address this, I construct summary indices that are robust to multiple inference. That is, the probability of a false rejection (Type I error) does not increase as additional outcomes are added. Additionally, combining multiple outcomes into a single index reduces measurement error by averaging across outcomes. Following Peter C. O'Brien (1984) and Jeffrey R. Kling, Jeffrey B. Liebman, and Lawrence F. Katz (2007), I normalize each outcome to have a mean of zero and a standard deviation of one. Next, I equalize signs across outcomes, so that positive values of the index represent "good" outcomes. Finally, I create a new summary index variable that is the simple average of all outcomes.[12]

I construct summary indices for the three test scores, an index of the nontest score school-age outcomes (grade retention and learning disability diagnosis), and an index of the six long-term outcomes. Later, the test score index is separated further into age categories to look at the initial effect of Head Start and fade-out over time.

B. *Selection Bias*

Because children in Head Start come from very disadvantaged families, a simple comparison to children in other preschools or no preschool on outcomes such as test scores or educational attainment will be biased downward. In the absence of a randomized experiment, the challenge is to counteract this downward bias using nonexperimental methods. This motivates the use of family fixed effects. Equation (1) captures the basic identification strategy employed in the remainder of the paper, and in previous analyses of Head Start (CT 1995; GTC 2002). The regression includes a set of pre-treatment covariates and a family fixed effect, which ensures that differences in important covariates such as permanent income and maternal AFQT, indeed any time-invariant factor, are differenced out of the regression. The identifying assumption is that selection into Head Start *among members of the same family*

[12] O'Brien (1984) recommends weighting by the inverse of the sample covariance matrix to account for dependence across outcomes. I do this in specifications not reported here, but it makes little difference, so I use the simple average instead, again following Kling, Liebman, and Katz (2007).

is uncorrelated with the unobservable determinants of outcomes. The estimating equation is

$$Y_{ij} = \alpha + \beta_1 HS_{ij} + \beta_2 PRE_{ij} + \delta \mathbf{X}_{ij} + \gamma_j + \varepsilon_i, \tag{1}$$

where i indexes individuals and j indexes the family, $\mathbf{X}$ is a vector of family-varying controls, and γ_j is the family fixed effect. HS_{ij} and PRE_{ij} are the estimated effect of Head Start and other preschools, respectively, on the outcomes $Y_{ij.}$ Threats to validity emerge from the child-specific error term ε_i; more specifically, if $E(\varepsilon_i | \mathbf{X}_{ij}, HS_{ij}, PRE_{ij}, \gamma_j) \neq 0$. In principle, nonrandom assignment of siblings to preschool status could bias the results in either direction. If, for example, parents could only send one child to preschool, they might compensate for existing disparities by sending the child of lower ability. This would bias the estimates downward by attenuating sibling differences in test score performance and later outcomes. Alternatively, the parent may engage in favoritism, and if that favoritism extends to unobserved child-specific investments, program estimates will be biased upward. Although assignment to Head Start clearly is not random in an experimental sense, if the variation is uncorrelated with the outcomes of interest, then estimates of the Head Start treatment effect will be unbiased.

Still, *something* is driving differences in participation among siblings. One optimistic possibility is that it is idiosyncratic availability around the age of eligibility. Head Start is perennially underfunded and oversubscribed, and administrative guidelines require that all Head Start centers keep a waiting list (Office of Head Start 2008). Furthermore, since Head Start is fully subsidized for eligible families, there is no direct budget constraint for intra-household enrollment decisions. Although enrollment data from this period is unavailable, I can examine rates of enrollment in the centers that participated in the Head Start National Impact Study (HSNIS) in 2002. The experimental design of the HSNIS was based on the principle that no eligible child could be denied coverage. Since the randomization was performed at the level of the center, only centers with more eligible children living in the area than they could serve were included in the study (HHS 2005). These areas, forced to deny service to eligible children, comprised 85 percent of the population of Head Start centers (HHS 2005). Furthermore, this percentage was higher in urban areas, which are disproportionately represented in the CNLSY data (Currie and Neidell 2007). Ideally, enrollment was based on idiosyncratic factors such as area cohort size. Unfortunately, I have no direct evidence of the reasons for differences in enrollment between siblings.

In many other nonexperimental analyses, selection bias is driven by decisions made by program participants. However, children were enrolled in Head Start by age three or four, and so any nonrandom assignment must be driven by parental decisions that were made prior to, or at the age of, eligibility.[13] I address nonrandom assignment by assembling a series of pre-treatment covariates that do vary between siblings. I compare Head Start participants to their siblings who were not in the

[13] Furthermore, because parents were asked about Head Start participation contemporaneously rather than retrospectively, there is little possibility for biased recall of program participation.

program across all these covariates, looking for a pattern that explains differential participation. I also control for these covariates directly in the estimation of the effect of the program.

Table 2 examines differences in pre-treatment covariates by preschool status for a wide array of covariates, including age, birth order, maternal work history, child-care arrangements, family structure, infant health, and income around the age of eligibility. Each row represents a separate regression of each covariate on Head Start and other preschool indicators, and a family fixed effect, with no other controls. Standard errors are in parentheses (clustered at the family level). I also report the control mean and standard deviation in the third column of each panel, and the last column reports the sample size for that particular covariate. Since the CNLSY started in 1986, some of the children of the NLSY women were already older than three when the survey began. Although some covariates are available dating back to 1979 through the regular NLSY, others begin in 1986 and are missing for older respondents. I account for missing data by imputing the mean value for the estimation sample, and I include a dummy variable for imputed responses in the outcome regressions in Section IV. The attrition rate in this sample was only about 4 percent, with no significant difference by preschool status. For the analyses that follow, in the remainder of the paper, the final sample size is 1,251.

Overall, there are no more differences in preschool status than what might be expected by chance variation. To provide a more formalized test of overall differences, I construct a summary index of pre-treament covariates according to the process outlined in Section IIIA. Where the appropriate "sign" of the outcome is unclear (such as whether it is "good" to be first-born, or to have a mother that works), I follow the gradient in the long-term summary index of outcomes. There is no significant difference between children in Head Start and children in no preschool or other preschools on this index, which is reported in the last row of Table 2. Although one possible reason is that these covariates are noisy and not predictive of future outcomes, I show that their inclusion in a regression of outcomes on preschool status increases the explanatory power of the regression substantially across families, but only modestly within them, and that it does so without changing the estimated effect of the program.

In the second row of Table 2, I exploit the timing of test administration in the CNLSY to construct a "pretest" for a subset of age-eligible children. Specifically, I examine the age three test scores of children, excluding those who were already enrolled in Head Start at age three. The PPVT, unlike the PIAT tests, is administered for the first time when the child is three years old. However, because the CNLSY is administered every two years, many children are age four or five when they take the test for the first time. Thus, the "pretest" sample is relatively small, and, to improve precision, I include controls for age (in months) and gender. Although the test has low power, there is no significant difference in age three test scores between subsequent Head Start enrollees and other children.

In the third and fourth rows of Table 2, I present results for the log of birth weight and an indicator for very low birth weight (less than 3.31 lbs). Head Start participants weigh about 4.8 percent more than their siblings who are not enrolled in preschool, and they are about 2 percentage points less likely to be very low birth weight. Given

TABLE 2—SIBLING DIFFERENCES IN PRE-TREATMENT COVARIATES, BY PRESCHOOL STATUS

	Head Start (1)	Other preschool (2)	Control mean (3)	Sample size (4)
Attrited	0.022	–0.008	0.038	1,314
	(0.013)	(0.016)	[0.192]	
PPVT score, age 3	2.24	–7.16*	19.90	195
	(4.82)	(4.12)	[11.10]	
ln (birth weight)	0.048**	–0.006	4.702	1,226
	(0.020)	(0.017)	[0.248]	
Very low BW (<3.31 lbs)	–0.022*	–0.004	0.021	1,226
	(0.012)	(0.008)	[0.145]	
In mother's HH, 0–3	0.002	–0.028	0.899	1,187
	(0.029)	(0.027)	[0.302]	
Pre-existing health limitation	–0.001	–0.041**	0.040	1,187
	(0.014)	(0.018)	[0.197]	
Firstborn	0.016	–0.124**	0.419	1,251
	(0.055)	(0.055)	[0.494]	
Male	0.000	–0.003	0.503	1,251
	(0.046)	(0.046)	[0.500]	
Age in 2004 (in years)	0.182	–0.433*	23.20	1,251
	(0.298)	(0.249)	[2.88]	
HOME score, age 3	1.98	3.07	38.05	427
	(3.24)	(4.10)	[26.25]	
Father in HH, 0–3	0.009	–0.003	0.624	739
	(0.034)	(0.023)	[0.450]	
Grandmother in HH, 0–3	–0.003	–0.049***	0.215	1,190
	(0.024)	(0.019)	[0.325]	
Maternal care, age 0–3	0.019	–0.015	0.689	1,244
	(0.019)	(0.022)	[0.405]	
Relative care, age 0–3	–0.007	0.022	0.180	1,244
	(0.019)	(0.019)	[0.335]	
Nonrelative care, age 0–3	–0.012	–0.006	0.131	1,244
	(0.017)	(0.016)	[0.283]	
Breastfed	–0.053**	–0.010	0.333	1,234
	(0.027)	(0.024)	[0.472]	
Regular doctor's visits, age 0–3	0.043	–0.055	0.383	430
	(0.102)	(0.110)	[0.488]	
Ever been to dentist, age 0–3	0.033	0.008	0.303	401
	(0.137)	(0.137)	[0.461]	
Weight change during pregnancy	0.056	–0.168	29.71	1,146
	(1.181)	(1.139)	[15.34]	
Child illness, age 0–1	0.016	–0.061	0.520	1,175
	(0.042)	(0.041)	[0.500]	
Premature birth	–0.048	0.007	0.218	1,175
	(0.034)	(0.034)	[0.413]	
Private health insurance, age 0–3	0.093	0.032	0.447	431
	(0.069)	(0.049)	[0.481]	
Medicaid, age 0–3	0.048	–0.006	0.376	431
	(0.060)	(0.043)	[0.456]	
ln (income), age 0–3	–0.012	0.043	9.99	1,186
	(0.043)	(0.033)	[0.72]	
ln (income), age 3	0.011	0.054	9.98	993
	(0.085)	(0.064)	[0.83]	
Mom average hours worked, year before birth	–1.11	2.06	26.03	377
	(3.14)	(1.87)	[12.15]	
Mom average hours worked, age 0–1	–1.08	1.77	32.52	379
	(3.16)	(1.72)	[11.07]	
Mom smoked before birth	–0.012	–0.005	0.392	1,186
	(0.030)	(0.023)	[0.489]	
Mom drank before birth	0.004	0.010	0.080	1,251
	(0.021)	(0.021)	[0.272]	
Pre-treatment index	0.014	0.047	–0.063	1,251
	(0.061)	(0.055)	[0.987]	

Notes: The first and second rows in each column heading of the table are the coefficients on Head Start and other preschools, respectively, from a regression with each pre-treatment variable as the outcome, and a family fixed effect. Standard errors are in parentheses and are clustered at the family level. The third row contains the control mean for each covariate with the standard deviation in brackets, and the fourth row is the sample size for that covariate. Responses are missing for some mothers that already had children age three and above by the first survey year. Other variables are present only in the Mother-Child supplement, which began in 1986.

***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

the emerging literature on the connection between birth weight and later outcomes (e.g., Sandra E. Black, Paul J. Devereux, and Kjell G. Salvanes 2007), this is a serious threat to the validity of the estimates. There are a few reasons to believe that birth weight differences are not a serious source of bias, however. First, it appears that the difference is caused by a disproportionate number of low-birth-weight children, rather than by a uniform rightward shift in the distribution of birth weight for Head Start children. For example, there are no significant differences in birth weight once low-birth-weight children (who represent less than 10 percent of the sample) are excluded.

Second, there is an important interaction between birth order and birth weight in this sample. Most of the difference in mean birth weight comes from children who are born third, fourth, or later. Later-birth-order children who subsequently enroll in Head Start are much less likely to be low birth weight than their older siblings who did not enroll in preschool. When I restrict the analysis to sibling pairs only, birth weight differences are much smaller and no longer significant, and the main results are unaffected. Finally, I estimate all the models in Section V with low-birth-weight children excluded, and, again, the main results are unchanged.

Still, to get a sense of the magnitude of any possible positive bias, I back out a correction using the long-run effect of birth weight on outcomes estimated by Black, Devereux, and Salvanes (2007). Specifically, they find that 10 percent higher birth weight leads to an increase in the probability of high school graduation of 0.9 percentage points for twins and 0.4 percentage points for siblings.[14] If that reduced form relationship holds here, a simple correction suggests that the effect of Head Start on high school graduation (and by extension, other outcomes) could be biased upward by between 0.2 and 0.4 percentage points, or about 2–5 percent of the total effect.

IV. Results

In Section III, I show that Head Start participants are negatively selected across families. If this is true, inclusion of covariates that are positively correlated with the outcomes of interest will increase the estimated effect of Head Start. I also assert that within-family differences in sibling participation are uncorrelated with long-term outcomes. To show these patterns directly, Table 3 presents results for the effect of Head Start on test scores, with an increasing number of covariates added to the regression. Each column estimates a form of equation (1), and includes controls for gender and first born status, plus age-at-test and year fixed effects. Rather than allowing the treatment effect to vary each year, I create three age categories—initial (age 5–6), primary school (age 7–10), and adolescent (age 11–14)—and I interact them with indicators for Head Start and other preschool programs.

Without any additional controls (column 1), Head Start participants score about 0.02 standard deviations lower at ages 5–6 than children who are not in preschool.

[14] These figures are taken from table III of Black, Devereux, and Salvanes (2007). Of course, the institutional context is different in their case (population data from Norway), and one could argue the effect of low birth weight may be greater for low-income children.

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

	(1)	(2)	(3)	(4)	(5)
Head Start					
Ages 5–6	–0.025	0.081	0.093	0.131	0.145*
	(0.091)	(0.083)	(0.079)	(0.087)	(0.085)
Ages 7–10	–0.116	0.040	0.067	0.116*	0.133**
	(0.072)	(0.065)	(0.061)	(0.060)	(0.060)
Ages 11–14	–0.201***	–0.053	–0.017	0.029	0.055
	(0.070)	(0.065)	(0.061)	(0.061)	(0.062)
Other preschools					
Ages 5–6	0.167**	0.022	–0.019	–0.102	–0.079
	(0.083)	(0.082)	(0.078)	(0.084)	(0.085)
Ages 7–10	0.230***	0.111*	0.087	0.031	0.048
	(0.070)	(0.064)	(0.061)	(0.061)	(0.065)
Ages 11–14	0.182**	0.076	0.037	–0.040	–0.022
	(0.072)	(0.068)	(0.065)	(0.066)	(0.069)
Permanent income (standardized) mean (0), SD (1)			0.112*		
			(0.064)		
Maternal AFQT (standardized) mean (0), SD (1)			0.353***		
			(0.057)		
Mom high school			0.141**		
			(0.071)		
Mom some college			0.280***		
			(0.080)		
p (all age effects equal—Head Start)	0.074	0.096	0.161	0.092	0.151
Pre-treatment covariates	N	Y	Y	N	Y
Sibling fixed effects	N	N	N	Y	Y
Total number of tests	4,687	4,687	4,687	4,687	4,687
R^2	0.028	0.194	0.268	0.608	0.619
Sample size	1,251	1,251	1,251	1,251	1,251

Notes: The outcome variable is a summary index of test scores that includes the child's standardized PPVT and PIAT math and reading scores at each age. Head Start and other preschool indicators are interacted with the three age groups (5–6, 7–10, and 11–14) listed above. Each column includes controls for gender, first born status, and age-at-test and year fixed effects, plus the covariates indicated in the bottom rows. The unit of observation is child-by-age. Standard errors are clustered at the family level.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

The point estimate decreases, over time, to about −0.20 standard deviations by ages 11–14. Children in other preschool programs score between 0.16 and 0.23 standard deviations higher without any additional controls. Column 2 includes all of the pre-treatment covariates in Table 2. The estimated effect of Head Start is now slightly positive initially (though not significant), but it becomes negative again by ages 11–14. The effect of other preschools is lower at each age group and no longer significant. Column 3 adds controls for three powerful indicators of socioeconomic status (SES): permanent income, maternal AFQT, and maternal education; and the estimated effect of Head Start continues to rise.

Column 4 includes a family fixed effect only, while column 5 adds all available covariates. If Head Start participants were positively selected within families, we

might expect the inclusion of pre-treatment covariates to reduce the estimated effect of Head Start. Instead, including them in the regression increases the coefficients slightly. Because these covariates are predictive of test scores even within families (the R^2 increases slightly from 0.608 to 0.619), the estimated effect of Head Start might be a bit higher if selection on unobservables were similar to selection on observables (Joseph G. Altonji, Todd E. Elder, and Christopher R. Taber 2005). Overall, the effect of Head Start on test scores in column 5 (the preferred specification) is 0.145 standard deviations at ages 5–6. This is roughly consistent with the results in CT (1995) and the Head Start National Impact Study (HHS 2005; Ludwig and Phillips 2008). I find some evidence of test score fade-out. The effect is 0.133 standard deviations for ages 7–10, and 0.055 standard deviations for ages 11–14. This fade out pattern is consistent with the results of many other interventions (CT 1995; Krueger and Whitmore 2001; Banerjee et al. 2007; Anderson 2008), but still I cannot reject the hypothesis that test score effects are equal across all age groups ($p = 0.151$). In general, the coefficient on Head Start increases and the coefficient on other preschools decreases as covariates are added to the regression. This pattern holds for other outcomes as well.

Panel A of Table 4 contains the main results of the paper. Each coefficient is from an estimation of equation (1) with all pre-treatment covariates, as in column 5 of Table 3. The first three columns come from the same regression as Table 3, with Head Start and other preschool treatment effects interacted with age groups. The fourth column combines them all into one age group, but the results are otherwise identical to column 5 of Table 3. The fifth column contains results for the school age nontest score outcomes: grade retention and learning disability diagnosis. Head Start participants score 0.265 standard deviations higher than their siblings who do not attend preschool. Other preschools also have an effect that is nearly as large (0.172 standard deviations). Column 6 contains results for the six-item index of young adult outcomes: high school graduation, college attendance, idleness, crime, teen parenthood, and health status. I estimate that participation in Head Start leads to a statistically significant impact of 0.228 standard deviations, relative to children who are not enrolled in preschool. This is a very large effect, equal to about a third of the difference in outcomes between the bottom quartile and the median permanent income, and about 75 percent of the black-white outcome gap in this sample. By contrast, an initial test score gain of 0.145 standard deviations closes about 25 percent of the permanent income gap and about 20 percent of the black-white gap. Taken together, the results suggest that the long-term impact of Head Start is larger than what would be predicted even by initial test score gains. If, instead, we consider "final" age 11–14 test score gains, then the long-term impact is much larger.

Panels B–D of Table 4 contain Head Start treatment effects for race, gender, and maternal AFQT subgroups. Like panel A, each regression coefficient comes from an estimation of equation (1) using the covariates in column 5 of Table 3, the preferred specification. Panel B shows results separately for black and white or Hispanic children. Initial test score gains are very large for African Americans (0.287 standard deviations), but they fade out to near 0 by ages 11–14. We can strongly reject the hypothesis that test scores are equal across age groups for

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

	Test scores				Nontest score	Long term
	5–6 (1)	7–10 (2)	11–14 (3)	5–14 (4)	7–14 (5)	19+ (6)
Panel A: Overall						
Head Start	0.145*	0.133**	0.055	0.101	0.265***	0.228***
	(0.085)	(0.060)	(0.062)	(0.057)	(0.082)	(0.072)
Other preschools	–0.079	0.048	–0.022	–0.012	0.172*	0.069
	(0.085)	(0.065)	(0.069)	(0.062)	(0.088)	(0.072)
p (HS = preschool)	0.021	0.254	0.315	0.118	0.372	0.080
Panel B: By race						
Head Start (black)	0.287***	0.127*	0.031	0.107	0.351***	0.237**
	(0.095)	(0.075)	(0.076)	(0.072)	(0.120)	(0.103)
Head Start (white/Hispanic)	–0.057	0.111	0.156	0.110	0.177	0.224**
	(0.120)	(0.092)	(0.095)	(0.090)	(0.111)	(0.102)
p (black = nonblack)	0.024	0.895	0.308	0.982	0.282	0.924
Panel C: By gender						
Head Start (male)	0.154	0.181**	0.141**	0.159**	0.390***	0.182*
	(0.107)	(0.079)	(0.081)	(0.076)	(0.123)	(0.103)
Head Start (female)	0.128	0.059	0.033	0.055	0.146	0.272**
	(0.106)	(0.083)	(0.085)	(0.081)	(0.108)	(0.106)
p (male = female)	0.862	0.287	0.357	0.346	0.135	0.553
Panel D: By maternal AFQT score						
Head Start (AFQT $\leq$ –1)	0.171	0.016	–0.023	0.015	0.529***	0.279**
(n = 361)	(0.129)	(0.095)	(0.102)	(0.094)	(0.156)	(0.114)
Head Start (AFQT > –1)	0.133	0.172**	0.144*	0.154**	0.124	0.202**
(n = 890)	(0.094)	(0.073)	(0.074)	(0.071)	(0.091)	(0.091)
p (low = high AFQT)	0.809	0.198	0.192	0.245	0.024	0.595
Panel E: P-values for equality of test scores by age group						
	Black	Nonblack	Male	Female	Low AFQT	High AFQT
p (all effects equal)	0.003	0.240	0.262	0.254	0.198	0.205

Notes: All results are reported using the specification in column 5 of Table 3, which includes a family fixed effect, all pre-treatment covariates, and controls for gender, age, and firstborn status. Race and gender subgroup estimates are obtained by interacting the Head Start treatment effect with a full set of dummy variables for each subgroup. Standard errors are in parentheses and are clustered at the family level. The test score indices include the PPVT and PIAT Math and Reading Recognition tests. The nontest score index includes indicator variables for grade retention and learning disability diagnosis. The long-term outcome index includes high school graduation, college attendance, idleness, crime, teen parenthood, and self-reported health status.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

African Americans ($p = 0.003$). By comparison, the test score gains for white and Hispanic children do not fade out, but are actually slightly negative initially and then increase to 0.156 standard deviations for ages 11–14. Pooled age 5–14 results by race are similar despite the different pattern by age group. Nontest score gains are larger for African Americans (0.351 versus 0.177 standard deviations) but long-term gains are similar by race.

Panel C presents results by gender, and here we see that test score improvements occur almost entirely among male children. The initial gain is similar (0.154 standard deviations for males versus 0.128 for females), but male children experience less test score fade-out. Finally, in panel D, I split the sample into children whose mothers have "low" (defined as one standard deviation below the mean for the full NLSY sample) and "high" AFQT scores. Initial test score gains are similar, but children with low AFQT mothers experience near complete fade out by ages 7–10 (0.016 standard deviations), whereas children of high AFQT mothers maintain their gains through ages 11–14 (0.144 standard deviations versus −0.023 for the low AFQT sample). Still, non-test score gains are significantly larger for the low AFQT sample (0.524 versus 0.124, $p = 0.024$), and long-term gains are modestly higher (0.279 versus 0.202) despite the absence of test score gains after age 6.

In no case can we reject the hypothesis that long-term gains are equal across subgroups. Still, the case for projections of long-term impacts based on initial or overall test score increases is weak. This is particularly true for the children of low AFQT mothers, for whom there is no test score gain despite a very large long-term impact. Finally, long-term gains are larger for African Americans despite test score fade out. This matches evidence from model preschool programs, which was based on entirely African American samples, and is consistent with the notion that test score gains are an incomplete measure of long-term benefits.

A. *Individual Outcomes*

Table 5 presents results from regressions of the form in equation (1), where the outcomes are individual variables in the noncognitive and long-term summary indices. I report point estimates and standard errors for the overall sample and by race, gender, and maternal AFQT subgroups. Head Start participants are about 8.5 percentage points more likely to graduate from high school, 6 percentage points more likely to have attempted at least one year of college, 7 percentage points less likely to be idle, and 7 percentage points less likely to be in poor health. A few patterns are notable. First, excluding the GED from the high school graduation variable reduces the coefficient somewhat.[15] Second, the pattern of impacts is quite different by race and gender. The overall increase in college attendance and health status is driven largely by females, whereas the results for "idleness" and grade retention hold mostly among males. Gains in educational outcomes such as grade retention, learning disability diagnosis, high school graduation, and college attendance are much larger for African Americans. Finally, note the large increase in high school graduation (nearly 17 percentage points) for the children of low AFQT mothers. In contrast, only high maternal AFQT children are more likely to attend college (about 8 percentage points, compared to 1 percentage point for the low AFQT sample).

[15] All summary index calculations include GED recipients as graduates. Excluding GED recipients reduces the value of the index slightly, but it remains statistically significant.

TABLE 5—POINT ESTIMATES FOR INDIVIDUAL OUTCOMES

	All	Black	Nonblack	Male	Female	Low AFQT	High AFQT
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Grade repetition	–0.069*	–0.107*	–0.027	–0.204***	0.055	–0.140**	–0.031
	(0.040)	(0.056)	(0.059)	(0.058)	(0.057)	(0.069)	(0.050)
Learning disability	–0.059***	–0.071**	–0.046	–0.047	–0.070***	–0.109***	–0.032
	(0.021)	(0.028)	(0.030)	(0.030)	(0.026)	(0.042)	(0.021)
High school graduation	0.086***	0.111***	0.055	0.114**	0.058	0.167***	0.042
	(0.031)	(0.041)	(0.048)	(0.048)	(0.044)	(0.056)	(0.036)
not including GED	0.063*	0.067	0.058	0.108**	0.021	0.126**	0.027
	(0.034)	(0.044)	(0.051)	(0.052)	(0.047)	(0.063)	(0.038)
At least one year of college attempted	0.057	0.136***	–0.034	0.022	0.091*	0.012	0.082*
	(0.036)	(0.049)	(0.050)	(0.045)	(0.054)	(0.051)	(0.047)
Idle	–0.071*	–0.030	–0.123**	–0.100**	–0.043	–0.070	–0.072
	(0.038)	(0.053)	(0.055)	(0.049)	(0.052)	(0.070)	(0.045)
Crime	0.019	0.051	–0.020	0.036	0.002	0.038	0.008
	(0.040)	(0.050)	(0.062)	(0.058)	(0.057)	(0.072)	(0.047)
Teen parenthood	–0.019	–0.040	–0.001	0.011	–0.047	–0.038	–0.008
	(0.036)	(0.052)	(0.053)	(0.052)	(0.056)	(0.065)	(0.043)
Poor health	–0.070***	–0.047	–0.094**	–0.036	–0.102**	–0.090*	–0.060*
	(0.026)	(0.035)	(0.043)	(0.037)	(0.042)	(0.047)	(0.033)

Notes: Results for each outcome are reported using the specification in column 5 of Table 3, which includes a family fixed effect, all pre-treatment covariates, and controls for gender, age, and firstborn status. Race and gender subgroup estimates are obtained by interacting the Head Start treatment effect with a full set of dummy variables for each subgroup. Standard errors are in parentheses and are clustered at the family level.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

B. *Robustness Checks*

I probe the robustness of the results in several ways. First, I experiment with alternative sample selection rules, such as eliminating children from the Head Start sample who reported participation inconsistently, or by eliminating the small number of children who indicated less than three months of participation in the program. The age restriction (age 19 or above by 2004) was designed to ensure that respondents were no longer in school while maximizing the size of the sample. Since some outcomes may be particularly sensitive to age, I reestimate the model using age 20 and above instead. A related concern is that the results are driven by differences between siblings that are very different in age, so I restrict the analysis to siblings who are no more than five years apart. I also cap the age of the sample at 28 years and then at 25 years, dropping respondents for whom there is no pre-treatment data and for whom participation in Head Start begins before the first CNLSY survey year, respectively. I also restrict the analysis to sibling pairs only. Finally, I reestimate the treatment effects in Tables 4 and 5 with the other preschool dummy variable excluded, to compare Head Start with what sample members would have received in expectation in

the absence of the program. Although individual point estimates are, at times, sensitive to these rules, none of them changes the qualitative nature of the findings or the statistical significance of the long-term summary index.

The nature of the identification strategy might also attenuate the estimated effect of Head Start if there are treatment spillovers between children. This is particularly likely given Head Start's focus on parenting practices (Office of Head Start 2008). GTC (2002) test for treatment spillovers by interacting the Head Start treatment effect with an indicator for first born status. If there are spillovers from the older sibling to the younger, then Head Start will appear to have a larger impact on non-first-born children. In results not reported here, I find very limited and inconsistent evidence of spillovers.

A final issue concerns the weighting of the sample. The NLSY contains year-specific weights that are intended to make the sample nationally representative. The restriction to families with multiple siblings with varying preschool participation makes the sample clearly nonrepresentative, however. In results not reported here, I reestimate all models using a family-specific weight that is the simple average of each sibling's sample weight when they were four years old. The results, which can be thought of as nationally representative conditional on sample restrictions, are nearly unchanged.[16]

V. Comparability and Benefit-Cost Analysis

One way to benchmark the magnitude of the impact of Head Start is to compare it directly to "model" programs. Anderson (2008) performs a careful reanalysis of three early childhood interventions (Perry Preschool, Carolina Abecedarian, and the Early Training Project). His summary index of adult outcomes is similar in construction and variable usage, although he stratifies treatment effects by gender (and all of the children in the three studies were African American). If I simply compute an average of the treatment effects that is weighted by sample size, the overall impact of the three programs on adult outcomes is about 0.115 standard deviations, about half of the effect size here.[17] We might sensibly exclude the Early Training Project, however, which was the weakest intervention of the three and found essentially no long-term impact. In that case the average impact of Perry Preschool and Abecedarian was about 0.28 standard deviations for teenage outcomes and 0.26 standard deviations for adults, compared to the Head Start treatment effect of about 0.23 standard deviations that is estimated here.[18]

The summary index of long-term outcomes was constructed based on comparability to previous analyses and a priori notions of importance. I made no assumptions about the relative importance of different outcomes. But we might think, for example, that educational attainment is of particular value given the established

[16] The summary index value is 0.216 with a standard error of 0.082. Individual outcome results are available from the author upon request.

[17] This is taken from table 3 of Anderson (2008).

[18] The high estimated benefit-cost ratio for the Perry Preschool Project is based on large reductions in crime, however, and is an estimate of the marginal social benefit, whereas these results, as well as Anderson (2008), consider only the private return.

causal connection between education and outcomes such as earnings, crime, and health (Joshua D. Angrist and Alan B. Krueger 1991; Lance Lochner and Enrico Moretti 2004; David M. Cutler and Adriana Lleras-Muney 2006).

One approach to the accounting of the returns to Head Start might be to take increases in educational attainment as a proxy for other outcomes of interest, using benchmarks from the existing literature. Levin et al. (2007) estimate the social cost of high school dropout at about $256,000 using this methodology. One of the programs they evaluate is the Perry Preschool Project, which generated a 19 percentage point increase in high school graduation at a cost of $12,532 per student, for a per-graduate cost of $65,959. Using similar methodology, the per-graduate cost of Head Start is estimated to be $65,116.[19] Thus, according to this methodology the benefit-cost ratios for Head Start and Perry Preschool are quite similar. Finally, a speculative benefit-cost analysis based on a projection of adult wages from the NLSY-79 suggests that the "break-even" effect size for Head Start is only about 0.06 (0.12) standard deviations at a discount rate of 3 (5) percent. These projections are in the Appendix.

Belfield et al. (2006) find that two-thirds of the social return to investment in Perry Preschool comes from reductions in criminal activity. Since I find no impact of Head Start on crime here, this is potentially an important difference between the two programs. Even if the estimated overall effect size of Head Start is 80 percent of model programs, the translation to private or social benefits might not be one-to-one since outcomes such as criminal activity have an extraordinarily high social cost. It is worth pointing out, however, that the reduction in crime found in the Perry Preschool analysis was small at age 27 but much larger by age 40 (Barnett 1995; Belfield et al. 2006). Furthermore, previous research has found that self-reported crime data (unlike the arrest records used in the Perry Preschool study) are highly unreliable, both in the NLSY and in other data sources (Lochner and Moretti 2004; Kling, Ludwig, and Katz et al. 2005). In sum, we must exercise caution in comparing the benefit-cost ratios of programs that were evaluated differently.

VI. Discussion and Conclusion

This paper provides evidence of the long-term benefits of Head Start for a recent birth cohort of children. While my results rely on nonexperimental comparisons between siblings who differ in their participation in the program, I find little evidence of systematic within-family bias in preschool assignment, and the results are robust to sensitivity checks and alternative specifications. I estimate that the long-term impact of Head Start is about 0.23 standard deviations on a summary index of young adult outcomes, with larger impacts for African Americans and relatively disadvantaged children. This gain is about one-third of the size of the outcome gap between the bottom

[19] The estimate in Levin et al. (2007) was calculated net of reductions in special education and grade retention, which deflated the cost estimates in Barnett et al. (2007) by approximately 20 percent. I applied the same discount factor here. The estimated increase in high school graduation of 8.6 percentage points yields (100/8.6) $\times$ ($7000 $\times$ 0.8) = $65,116. My thanks to an anonymous referee for suggesting this calculation.

quartile and the median permanent income in the CNLSY sample, and is about 80 percent as large as the gains from the Perry Preschool and Carolina Abecedarian model preschool programs (Carneiro and Heckman 2003; Anderson 2008).

I find an initial age 5–6 test score gain of about 0.15 standard deviations that fades out to about 0.05 by ages 11–14. Fade-out is particularly strong for African Americans and for very disadvantaged children, and yet they experience the largest long-term gains. This does not rule out, for example, an increase in latent cognitive skills that is more poorly measured by the same test as children age—but it does imply that a projection of future benefits for these children based on test score gains alone would greatly understate the impact of the program.

In 2002, the US Department of Health and Human Services commissioned a large-scale randomized trial of Head Start. The first-year follow-up found "small" gains in test scores of between 0.1 and 0.2 standard deviations, which some have suggested might be a reasonable proxy for the long-term benefits of Head Start (HHS 2005; Ludwig and Phillips 2008). The results presented here, while not conclusive, suggest that such a one-to-one projection may be a lower bound for the total effect of the program.

Appendix: Benefit-Cost Analysis of Head Start Using the NLSY-79 Sample

A full accounting of the benefits of Head Start would measure (among other things) the increase in lifetime wages associated with participation in the program. While the participants in this survey are still too young for us to observe their adult wages, we can project future wage gains using the previous survey generation, the NLSY 1979. Since the NLSY was administered to youths aged 14–22, we can collect the same set of age 19 outcomes for this older cohort, and measure the relative contribution of each to adult wages. Intuitively, this projection exercise will be accurate only to the extent that the relative and absolute contributions of each outcome are constant across generations; and the marginal contribution of Head Start is equivalent to cross-sectional increases in each outcome. The true effect will eventually be known with additional years of data. Still, the results of this somewhat speculative exercise may be informative.

To project the impact of Head Start on wages, I first take all original members of the NLSY that are age 19 or less at the first date of the survey. For these individuals, I assemble a list of age 19 outcomes that is analogous to the outcomes in the long-term summary index.[20] Next, I assemble and average respondents' labor market wages, from age 20 until the last available survey year in 2004, when they are between 39 and 44 years old. I then estimate

$$\ln\left(Wages^{20-44}{}_i\right) = \beta \mathbf{S}_i + \delta \mathbf{X}_i + \varepsilon_i, \tag{2}$$

[20] All variables are exactly the same as in the CNLSY, except for health status, which was unavailable in the NLSY. In its place, I use the existence of a self-reported "limiting health condition."

TABLE A1—THE PROJECTED EFFECT OF HEAD START ON ADULT WAGES

Log of average wages, age 20+	(1)	(2)	(3)	(4)
Summary index			0.524	0.423
			(0.015)	(0.016)
High school grad	0.273	0.167		
	(0.014)	(0.016)		
Some college	0.147	0.099		
	(0.011)	(0.012)		
Idle	0.303	0.292		
	(0.017)	(0.017)		
Crime	0.070	0.071		
	(0.014)	(0.014)		
Teen pregnancy	0.035	0.027		
	(0.014)	(0.014)		
Health	0.044	0.042		
	(0.014)	(0.014)		
AFQT score		0.280		0.263
		(0.018)		(0.017)
AFQT squared		–0.102		–0.099
		(0.012)		(0.011)
R^2	0.260	0.284	0.260	0.283
Sample size	7,778	7,778	7,778	7,778

Benefit-cost analysis based on projected wage gains

	All	Low AFQT	High AFQT	Black	Nonblack	Male	Female
Coefficient on Head Start (weighted)	0.248	0.328	0.211	0.259	0.249	0.244	0.254
Coefficient from wage regression	0.423	0.607	0.351	0.442	0.405	0.327	0.487
Predicted mean log wage	9.52	8.81	9.71	9.29	9.57	9.84	9.19
Yearly wage gain	$1,507	$1,476	$1,267	$1,313	$1,520	$1,559	$1,290
Cost of program (in 2007 dollars)	$6,000						
Internal rate of return	7.9%	7.8%	7.2%	7.3%	7.9%	8.0%	7.2%

Notes: The dependent variable is the log of average wages for all years for which data are available in the NLSY 1979, starting at age 20 and ending in the last survey year (2004), which is between ages 39 and 44, depending on the respondent's initial age. Each of the independent variables is standardized to have a mean of zero and a standard deviation of one. The summary index variable is a composite of the six items listed below it, and the coefficients in column 2, and used as weights for the index in columns 3 and 4. The regression also includes age dummies, and standard errors are robust. The top row of the bottom panel is the coefficient on Head Start from a regression like column 6 of Table 7, except the summary index is weighted using the coefficients in column 2 of this table. The second row is the coefficient on the summary index in column 4 of this table. The third row calculates the mean log wage for NLSY 1979 respondents with the same average value of the summary index as Head Start participants in the CNLSY. The yearly wage gain is calculated by multiplying the coefficients in the first two rows together, adding that value to the predicted log wage, and subtracting the original prediction. For example, $0.248 \times 0.423 = 0.105$. $\exp(9.52 + 0.105) - \exp(9.52) = \$1{,}507$. The cost of the program (in adjusted 2007 dollars) is obtained from the HHS Web site, and is a weighted average for the years in which CNLSY respondents were enrolled. The internal rate of return is calculated assuming 1 year of enrollment at age 4, and that the yearly wage gain begins at age 20 and goes through age 65. This calculation does not include any social benefit or cost savings, and considers only private benefits that are directly capitalized into wages.

where the **X** vector contains race, gender, and age dummy variables and a quadratic in the respondents' standardized and age-normalized AFQT score, following Derek A. Neal and William R. Johnson (1996). The coefficients on each (standardized) outcome in the **S** vector give the marginal contribution of each to future wages, and

can be used as weights for the summary index. Using the estimated coefficients in this fashion, I generate a replica weighted summary index for the NLSY sample. This is a variant of the procedure for interpretation of regressions with multiple proxies outlined in Darren Lubotsky and Martin Wittenberg (2006). In column 4 of Appendix Table A1, we see that a one standard deviation increase in this summary index raises average yearly wages by 0.423 log points. To forecast the effect of Head Start on future wages, I reestimate the specification in column 6 of Table 4 with the weights obtained above, and I multiply this coefficient by the wage regression coefficient. For example, I estimate that Head Start improves outcomes by 0.248 standard deviations on this weighted index, and since a one standard deviation increase raises wages by 0.423 log points, the projected impact of Head Start on wages is $0.248 \times 0.423 = 0.105$ log points.

Finally, I apply this increase to the mean wage for NLSY respondents with the same average characteristics on the summary index as Head Start participants. This generates an estimated yearly wage gain that, appropriately discounted and compared to the cost of the program, yields a rough estimate of the net present value of investment in Head Start. Results of this calculation are in Appendix Table A1. I obtain cost estimates using a weighted average of the estimated (real) cost of the program for the years in which children were enrolled. The average cost of one year of the program in these years is about $6,000. Under these assumptions, the internal rate of return is 7.9 percent. Solving for a break-even effect size yields minimum effects of 0.06, 0.12, and 0.20 standard deviations for discount rates of 3, 5, and 7 percent, respectively. I make the same calculations for the subgroups of interest defined in Table 5. Returns are relatively stable across subgroups, primarily because groups whose participants experience bigger marginal gains from Head Start have lower mean wages.

REFERENCES

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151–84.

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): 1481–95.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4): 979–1014.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, 122(3): 1235–64.

Barnett, W. Steven. 1992. "Benefits of Compensatory Preschool Education." *Journal of Human Resources*, 27(2): 279–312.

Barnett, W. Steven, Jason T. Hustedt, Allison H. Friedman, Judi Stevenson Boyd, and Pat Ainsworth. 2007. *The State of Preschool 2007: State Preschool Yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

Belfield, Clive R., Milagros Nores, Steven Barnett, and Lawrence Schweinhart. 2006. "The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Followup." *Journal of Human Resources*, 40(1): 162–90.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2007. "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes." *Quarterly Journal of Economics*, 122(1): 409–39.

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.

Carneiro, Pedro M., and James J. Heckman. 2003. "Human Capital Policy." In *Inequality in America: What Role for Human Capital Policies?* ed. Benjamin Friedman, 77–240. Cambridge, MA: MIT Press.

Case, Anne, Darren Lubotsky, and Christina Paxson. 2002. "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review*, 92(5): 1308–34.

Cunha, Flavio, and James J. Heckman. 2007. "The Technology of Skill Formation." *American Economic Review*, 97(2): 31–47.

Currie, Janet. 2001. "Early Childhood Education Programs." *Journal of Economic Perspectives*, 15(2): 213–38.

Currie, Janet, and Matthew Neidell. 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, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review*, 85(3): 341–64.

Currie, Janet, and Duncan Thomas. 2000. "School Quality and the Longer-Term Effects of Head Start." *Journal of Human Resources*, 35(4): 755–74.

Cutler, David M., and Adriana Lleras-Muney. 2006. "Education and Health: Evaluating Theories and Evidence." National Bureau of Economic Research Working Paper 12352.

Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-Term Effects of Head Start." *American Economic Review*, 92(4): 999–1012.

Gormley, William T., Jr., and Ted Gayer. 2005. "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program." *Journal of Human Resources*, 40(3): 533–58.

Haskins, Ron. 2004. "Competing Visions." *Education Next*, 4(1): 26–33.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science*, 312(5782): 1900–1902.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review*, 91(2):145–49.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics*, 24(3): 411–82.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz. 2005. "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment." *Quarterly Journal of Economics*, 120(1): 87–130.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2004. "Incentives to Learn." National Bureau of Economic Research Working Paper 10971.

Krueger, Alan B. 2003. "Inequality: Too Much of a Good Thing" In *Inequality in America: What Role for Human Capital Policies?* ed. Benjamin Friedman, 1–77. Cambridge, MA: MIT Press.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Tests Results: Evidence from Project STAR." *Economic Journal*, 111(468): 1–28.

Levin, Henry M., Clive Belfield, Peter Muennig, and Cecilia Rouse. 2007. "The Public Returns to Public Educational Investments in African-American Males." *Economics of Education Review*, 26(6): 699–708.

Lochner, Lance, and Enrico Moretti. 2004. "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." *American Economic Review*, 94(1): 155–89.

Lubotsky, Darren, and Martin Wittenberg. 2006. "Interpretation of Regressions with Multiple Proxies." *Review of Economics and Statistics*, 88(3): 549–62.

Ludwig, Jens, and Douglas L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics*, 122(1): 159–208.

Ludwig, Jens, and Deborah A. Phillips. 2007. "The Benefits and Costs of Head Start." National Bureau of Economic Research Working Paper 12973.

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: 257–68.

Lyon, G. Reid. 1996. "Learning Disabilities." *The Future of Children*, 6(1): 54–76.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy*, 104(5): 869–95.

O'Brien, Peter C. 1984. "Procedures for Comparing Samples with Multiple Endpoints." *Biometrics*, 40(4): 1079–87.

Office of Head Start. 2008. United States Department of Health and Human Services. http://www.acf.hhs.gov/programs/ohs/about/fy2008.html. (accessed November 17, 2008).

Schweinhart, Lawrence, 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*. Ypsilanti, MI: High/Scope Press.

United States Department of Health and Human Services. 2001. *Head Start FACES: Longitudinal Findings on Program Performance, Third Progress Report.* Administration for Children and Families. Washington, DC, January.

United States Department of Health and Human Services. 2005. *Head Start Impact Study: First Year Findings.* Administration for Children and Families. Washington, DC, May.

United States Department of Labor. 2008. Bureau of Labor Statistics National Longitudinal Survey of Youth User Guide. http://www.nlsinfo.org/nlsy79/docs/79html/79text/front.htm. (accessed May 22, 2009).