Inputs in the Production of Early-childhood Human Capital: Evidence from Head Start

Christopher Walters*

February 27, 2014

Abstract

Studies of small-scale "model" early-childhood programs show that high-quality preschool can have transformative effects on human capital and economic outcomes. Evidence on the Head Start program is more mixed. Inputs and practices vary widely across Head Start centers, however, and little is known about variation in effectiveness within Head Start. This paper uses data from a multi-site randomized evaluation to quantify and explain variation in effectiveness across Head Start childcare centers. I answer two questions: (1) Is there meaningful variation in short-run effectiveness across Head Start programs? and (2) Is variation in Head Start effectiveness related to observed inputs? To answer the first question, I develop an empirical Bayes instrumental variables procedure that measures variation in local average treatment effects (LATE), accounting for non-compliance with experimental assignments. I estimate that the cross-center standard deviation of cognitive effects is 0.3 test score standard deviations, which is substantially larger than typical estimates of variation in teacher or school effectiveness. Next, I assess the role of inputs in generating this variation, focusing on inputs commonly cited as drivers of the success of small-scale model programs. My results show that Head Start centers offering full-day service boost cognitive skills more than other centers, while Head Start centers offering frequent home visiting are especially effective at raising non-cognitive skills. Other key inputs, including the High/Scope curriculum, teacher education and certification, and class size, are not associated with increased effectiveness in Head Start. Together, observed inputs explain a small share of the variation in Head Start effectiveness. These findings suggest that replicating the effects of successful programs may be difficult, as the factors responsible for their success are largely unidentified.

*University of California, Berkeley Economics (e-mail: crwalters@econ.berkeley.edu). I am grateful to Joshua Angrist, Aviva Aron-Dine, David Autor, David Chan, Hilary Hoynes, Guido Imbens, Patrick Kline, Christopher Palmer, Parag Pathak, Jesse Rothstein, Tyler Williams, and participants at the MIT labor economics field lunch for useful comments and suggestions.

1 Introduction

Studies of small-scale "model" early-childhood education programs show that preschool attendance can boost outcomes in the short- and long-run. In the High/Scope Perry Preschool Project, a randomized trial that took place in the early 1960s, 123 disadvantaged children were randomly assigned to either an intensive preschool program or a control group without access to the program. Subsequent analyses showed that participation in the Perry program increased average IQ at age 5 by nearly a full standard deviation, and had lasting impacts on educational attainment, criminal behavior, drug use, employment, and earnings (Anderson 2008; Berruta-Clement et al. 1984; Heckman et al. 2010b; Schweinhart et al. 1997, 2005).[1] Heckman et al. (2010a) estimate the annual social rate of return to the Perry Project at between 7 and 10 percent. The North Carolina Abecedarian Project, another small-scale intervention, had similarly dramatic effects (Campbell and Ramey 1994, 1995). The striking success of these programs has led some analysts to argue that the returns to educational intervention peak early in life (Heckman 2011). These findings have also motivated recent calls for expansion of publicly-provided preschool (Obama 2013).

In contrast, evidence on the effects of large-scale early childhood programs is more mixed. Early studies of Head Start, the largest early childhood program in the United States, showed positive effects on cognitive skills, child mortality, and long-term outcomes (Currie and Thomas 1995; Ludwig and Miller 2007; Garces et al. 2002; Deming 2009).[2] More recently, results from the Head Start Impact Study (HSIS), the first randomized evaluation of Head Start, show smaller, less-persistent gains. The HSIS experiment involved random assignment of more than 4,000 children to Head Start or a control group at over 300 childcare centers throughout the US. The HSIS treatment group outscored the control group by roughly 0.1 standard deviations on measures of cognitive skill during preschool, but these gains did not persist into kindergarten (US Department of Health and Human Services 2010, 2012; Bitler et al. 2012). Moreover, the HSIS experiment showed little evidence of effects for a wide range of non-cognitive and health outcomes (US Department of Health and Human Services 2010).[3]

Inputs and practices vary widely across Head Start centers, however, and little is known about variation in effectiveness within Head Start. This paper uses HSIS data to quantify and explain variation in effectiveness across Head Start childcare centers. I focus on one candidate explanation for differences in the effects of early childhood programs: differences in inputs and practices used by these programs. Some Head Start centers use inputs more similar to successful model programs than others. For example, one-third of Head Start centers use the High/Scope curriculum, the centerpiece of the Perry Preschool experiment. Head Start centers also differ with respect to teacher characteristics, class size, instructional time, and frequency of

[1] Anderson (2008) argues that the Perry Project produced significant long-term benefits only for girls.

[2] Other studies finding positive effects of larger-scale programs include analyses of the Chicago Child-Parent centers and some state pre-kindergarten programs (Reynolds 1998; Gormley and Gayer 2005; Wong et al. 2008). Cascio and Schanzenbach (2013) find small effects of programs in Georgia and Oklahoma for poor children, and no effects for richer children. Fitzpatrick (2008) finds small effects for Georgia's program, though some subgroups benefit.

[3] In other analyses of the HSIS data, Gelber and Isen (2013) show that Head Start participation increaseds parental involvement with children after the program ended, while Bitler et al. (2012) show that the reduced form impact of assignment to Head Start is larger for lower quantiles of the distribution of Peabody Picture and Vocabulary Test (PPVT) scores.

home visits, all of which have been cited as central to the success of model programs (Schweinhart 2007). The aim of this paper is to assess the contribution of these key inputs to cross-center differences in Head Start effects.

My analysis proceeds in two steps. First, to ask whether there is meaningful variation to be explained by inputs, I quantify heterogeneity in causal effects across Head Start centers. This investigation is complicated by non-compliance with random assignment in the HSIS experiment. Instrumental variables (IV) is the standard procedure for dealing with non-compliance, but IV has poor properties in samples the size of the applicant pools at individual Head Start centers (Nelson and Startz 1990). As a consequence, the conventional empirical Bayes "shrinkage" approach to quantifying effect heterogeneity, which requires unbaised estimates with known sampling variance, is invalid when applied to center-specific IV estimates.

I therefore develop an empirical Bayes instrumental variables procedure which allows me to measure cross-center variation in compliance probabilities and local average treatment effects (LATE; Imbens and Angrist 1994). This procedure can be applied in more general settings to quantify heterogeneity in instrumental variables estimands across many groups. Estimates using this method reveal substantial heterogeneity in short-run Head Start effectiveness: the cross-center standard deviation of short-run cognitive effects is 0.3 test score standard deviations, which is substantially larger than recent estimates of variation in teacher and school effectiveness (Deming 2014; Chetty et al. 2013a; Kane et al. 2008).

In a second step, I ask whether this variation can be explained by differences in observed inputs. My results show that inputs explain little of the variation in Head Start effectiveness, and most inputs thought to be important to the success of model programs are uncorrelated with effectiveness in Head Start. Some inputs do play a role: Head Start centers offering full-day programs boost cognitive skills more than other centers, while centers offering frequent home visits are especially effective at raising non-cognitive skills. High/Scope Head Start centers are no more effective than other centers, however, and short-run effectiveness is uncorrelated with teacher education, teacher certification, and class size. Together, observed inputs explain less than 15 percent of the variation in Head Start cognitive effects. These findings suggests that replicating the effects of successful programs may be difficult, as the factors responsible for their success are largely unidentified.

The rest of the paper is organized as follows. The next section provides background on Head Start and describes the HSIS data. Section 3 summarizes the average impact of Head Start on summary indices of cognitive and non-cognitive skills. Section 4 outlines the empirical Bayes instrumental variables procedure used to investigate effect heterogeneity, and reports the results of this investigation. Section 5 analyzes the relationship between observed inputs and Head Start effectiveness. Section 6 concludes.

2 Data and Background

2.1 Head Start and the Head Start Impact Study

Head Start, the largest early-childhood program in the United States, enrolls roughly one million 3- and 4-year-old children at a cost of about $8 billion annually. Head Start awards grants to public, private non-profit, and for-profit organizations that provide childcare services to children below 130 percent of the Federal Poverty Line, though up to 35 percent of children attending a Head Start childcare center can be from households above this income threshold. Grantees are required to match at least 20 percent of federal Head Start funding. Head Start is based on a "whole child" model of school readiness that emphasizes non-cognitive social and emotional development in addition to cognitive skills. The grant-based nature of the program allows for a wide variety of childcare settings and practices, though all grantee agencies must meet a set of program-wide performance standards (US Department of Health and Human Services 2011; US Office of Head Start 2012).

The data used in my analysis come from the Head Start Impact Study (HSIS), a randomized evaluation of the Head Start program. The 1998 Head Start Reauthorization Act included a congressional mandate to determine the program's effects. As a result, the US Department of Health and Human Services (DHHS) conducted a nationally representative randomized controlled trial (DHHS 2010, 2012). The HSIS data includes information on 84 regional Head Start programs, 353 Head Start centers, and 4,442 children, each of whom applied to a sample Head Start center in Fall 2002. Sixty percent of applicants were randomly assigned the opportunity to attend Head Start ("treatment"), while the remaining applicants were denied this opportunity ("control"). Randomization took place at the Head Start center level; the HSIS data includes weights reflecting the probability of assignment for each child, which are used to adjust for these differences below.[4]

The HSIS sample includes two age groups, with 55 percent of students entering at age 3 and 45 percent entering at age 4. Three-year-old applicants could attend Head Start for up to two years before entering kindergarten, and three-year-olds assigned to the control group could re-apply to Head Start centers as four-year-olds the next year. Four-year-old applicants could attend for a maximum of one year. The data used here follow the treatment and control groups through 1st grade. DHHS (2010) provides a complete description of the HSIS experimental design and data collection procedures. The Online Appendix details the procedure used to construct my sample from the HSIS data.

2.2 Outcomes

The HSIS data include a large number of outcomes, collected for up to 4 years after random assignment. I organize these outcomes into summary indices of cognitive and non-cognitive skills. Table 1 lists the

[4]Some centers conducted multiple rounds of random assignment with differing admission probabilities, and the HSIS weights do not account for these differences. The discussion in DHHS (2010) suggests that any such differences are likely to be small, however.

outcomes included in each group. Cognitive outcomes include scores on the Peabody Picture and Vocabulary Test (PPVT) and several Woodcock Johnson III (WJIII) measures of cognitive ability. Non-cognitive outcomes, derived from parental surveys, include measures of social skills (making friends, hitting and fighting) and attention-span (concentration, restlessness). I exclude non-cognitive measures for which almost all respondents (90% or more) gave the same answer.[5]

Following Kling et al. (2007) and Deming (2009), I construct indices to summarize the impact of Head Start attendance across the outcomes listed in each column of Table 1. Specifically, I define the summary index

$$Y_i \equiv \frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{y_{i\ell}-\mu_\ell}{\sigma_\ell}\right),$$

where $y_{i\ell}$ is outcome ℓ for student i, and μ_ℓ and σ_ℓ are the control group mean and standard deviation of outcome ℓ. I define outcomes so that positive signs mean better performance, and standardize them separately by year and age cohort.

2.3 Applicant Characteristics

Head Start applicants typically come from families with low socioeconomic status. This can be seen in the first column of Table 2, which presents mean demographic characteristics for the HSIS control group. The demographic variables come from a baseline survey of parents conducted in the Fall of 2002; parents of 3,577 HSIS applicants (81 percent) responded to this survey. More than 35 percent of parents in the sample did not complete high school, and 18 percent of participants were born to teen mothers. The average household income in the sample is $1,290 per month.[6]

To check experimental balance, column (2) of Table 2 shows coefficients from regressions of baseline characteristics on assignment to Head Start, weighting by the HSIS baseline child weights to adjust for differences in the probability of assignment across centers. The treatment/control differences in means are statistically insignificant for all baseline variables except fathers' college attendance. The joint p-value from a test of the hypothesis that assignment to Head Start is unrelated to all baseline characteristics is 0.44, suggesting that random assignment to Head Start was successful.[7]

The last row of Table 2 shows that applicants assigned to Head Start were 66 percentage points more likely to participate in the program than applicants from the control group in the first year after random assignment. Twelve percent of students from the control group attended Head Start, most likely by applying

[5]The HSIS data also includes measures of non-cognitive skills reported by teachers. I do not use these measures since they are unavailable for many children before kindergarten, and my analysis focuses on outcomes during preschool.

[6]Some respondents did not report exact household incomes, but reported that their incomes fell within one of several ranges. For this calculation, I imputed income for these families as the midpoint of the reported range.

[7]Even with successful random assignment, non-random attrition has the potential to bias the experimental results. Appendix Table A1 shows attrition rates for the HSIS sample by year and outcome group, as well as treatment/control differences conditional on the controls included in Table 4. In preschool, outcomes are observed for 82 to 84 percent of children; the follow-up rate falls slightly in elementary school. Cognitive outcomes in preschool are observed slightly more frequently for children in the treatment group (3 percentage points) and this difference is marginally statistically significant. This modest differential attrition seems unlikely to drive the results reported below.

to other nearby Head Start centers outside the experimental sample. Twenty-two percent of children assigned to Head Start did not participate in the program. Together, these facts show that non-compliance with the experimental assignment is an important feature of the HSIS data, which motivates the instrumental variables approach taken below.

2.4 Center Characteristics

In addition to background information on applicants, the HSIS data includes detailed information on Head Start centers and their practices. I focus on inputs and practices that have been cited as central to the success of small-scale model programs. Schweinhart (2007) offers one view of the inputs that drove the success of the Perry Preschool Project:

> "The external validity or generalizability of the study findings extends to those programs that are reasonably similar to the High/Scope Perry Preschool Program. A reasonably similar program is a preschool education program run by teachers with bachelor's degrees and certification in education, each serving up to 8 children living in low-income families. The program runs 2 school years for children who are 3 and 4 years of age with daily classes of 2.5 hours or more, uses the High/Scope model or a similar participatory education approach, and has teachers visiting families at least every two weeks or scheduling regular parent events."

This account of the Perry program's effects emphasizes six key inputs: teacher education, teacher certification, class size, instruction time, the High/Scope curriculum, and home visiting. High/Scope is a participatory curriculum that emphasizes childrens' hands-on choices and experiences rather than adult-driven instruction (Epstein 2007). Schweinhart (2007) places particular weight on the High/Scope curriculum, arguing that results from the Perry Project and the followup High/Scope Preschool Curriculum Comparison Study "[suggest] that the curriculum had a lot to do with the findings."

No Head Start center replicates the Perry model, which used high levels of all six inputs and spent nearly three times as much as the average Head Start program on a per-pupil basis.[8] There is substantial variation in each of the six key Perry inputs within Head Start, however. This can be seen in Table 3, which summarizes characteristics of centers in the HSIS sample. One-third of Head Start centers use the High/Scope curriculum. Only 15 percent of Head Start staff members have bachelor's degrees, and 12 percent hold teaching licenses, but the fractions with these credentials range from zero to 100 percent across centers. The average Head Start center has 9.8 children for every staff member; the cross-center standard deviation of class size is 4 children. Sixty-two percent of Head Start centers provide full-day service, and 21 percent offer more than three home visits per year. In Section 5, I explore whether this variation in inputs can explain differences in effectiveness across Head Start centers.

[8] Heckman et al. (2010) report that the Perry program cost about $18,000 per child in 2006 dollars. Per-child expenditure in Head Start was $7,600 in 2011, which is $6,800 deflated to 2006 dollars using the Consumer Price Index series available at http://www.bls.gov (DHHS 2011).

3 LATE Framework and Pooled Estimates

3.1 LATE Framework

I next describe the local average treatment effect (LATE) framework used to interpret IV estimates of Head Start effects (Imbens and Angrist 1994). Consider a set of students, each applying to Head Start at one of J centers. Let $S_i \in \{1, 2, ..., J\}$ indicate the center where student i applies. The Bernoulli variable Z_i indicates random assignment to Head Start, and D_i indicates Head Start attendance. Let $D_i(1)$ and $D_i(0)$ indicate potential Head Start attendance as a function of Z_i, so that $D_i = D_i(Z_i)$. Finally, there is an outcome variable, Y_i, with potential values $Y_i(1)$ and $Y_i(0)$ indexed against D_i. The observed outcome is $Y_i = Y_i(D_i)$.

The LATE framework is built on the following assumptions:

1. Independence/Exclusion: $(Y_i(1), Y_i(0), D_i(1), D_i(0)) \perp\!\!\!\perp Z_i | S_i$
2. First Stage: $Pr[D_i(1) > D_i(0) | S_i = j] > 0 \ \forall j$
3. Monotonicity: $Pr[D_i(1) \geq D_i(0) | S_i = j] = 1 \ \forall j$

Assumption (1) follows if Z_i is randomly assigned within centers and only affects outcomes through Head Start attendance.[9] Assumption (2) requires that assignment to Head Start induces some children to attend Head Start at every center. Assumption (3) requires that random assignment to Head Start does not discourage any child from participating in the program.

Under these assumptions, the population can be partitioned into three groups defined by response to Head Start assignment. *Compliers* participate in Head Start if assigned to the program and not otherwise ($D_i(1) > D_i(0)$). *Never takers* decline to participate regardless of assignment ($D_i(1) = D_i(0) = 0$), and *always takers* participate even if assigned to the control group ($D_i(1) = D_i(0) = 1$). Imbens and Angrist (1994) show that instrumental variables estimates recover local average treatment effects (LATE), average causal effects of Head Start attendance for compliers. We have

$$\frac{E[Y_i | Z_i = 1, S_i = j] - E[Y_i | Z_i = 0, S_i = j]}{E[D_i | Z_i = 1, S_i = j] - E[Y_i | Z_i = 0, S_i = j]} = E[Y_i(1) - Y_i(0) | D_i(1) > D_i(0), S_i = j] \tag{1}$$

$$\equiv \beta^j.$$

β^j is the LATE at center j. The Wald (1940) instrumental variables estimator of β^j replaces population moments with sample moments on the left-hand side of equation (1).

[9] The exclusion restriction is not guaranteed by random assignment. One potential source of violation comes from control group students who attend Head Start centers outside the experimental sample. Specifically, the exclusion restriction is violated if the effectiveness of these non-experimental centers differs systematically from the effectiveness of experimental centers. The non-experimental centers attended by control group children are likely to be geographically close to the relevant experimental centers and administered by the same regional Head Start programs, so it seems plausible to assume that the effects of these centers are similar. Moreover, as discussed in Section 2, most non-compliance with experimental assignments involved treatment group members declining to attend Head Start, rather than control group members participating in the program.

3.2 Pooled Estimates

Before investigating heterogeneity in causal effects, I summarize the overall impact of Head Start using pooled equations of the form

$$Y_i = \alpha + \beta D_i + X_i'\lambda + \epsilon_i, \tag{2}$$

where X_i is a vector of the baseline controls from Table 2, included to increase precision. The attendance dummy D_i is instrumented with the Head Start assignment dummy Z_i, with first stage equation

$$D_i = \kappa + \pi Z_i + X_i'\delta + \eta_i. \tag{3}$$

I estimate these equations by weighted least squares using the HSIS baseline child weights to account for differences in the probability of assignment across centers. The coefficient β can be interpreted as a weighted average of center-specific LATEs.

Estimates of equations (2) and (3) reveal that Head Start attendance boosts outcomes during preschool, but these effects fade out immediately once children leave the program. Table 4 reports estimates of effects for cognitive and non-cognitive skills, separately by grade and assignment cohort. Column (1) shows that in the first year after random assignment, applicants assigned to treatment were 68 percentage points more likely to attend Head Start than applicants in the control group. The corresponding second-stage estimates for cognitive skills, reported in column (2), shows that Head Start attendance increases cognitive skills by 0.18 standard deviations for three-year-olds and 0.10 standard deviations for four-year-olds. These estimates are statistically significant at the 5-percent level, and (when appropriately adjusted for the first stage) they are consistent with intent-to-treat effects for individual outcomes reported by DHHS (2010) and Bitler et al. (2012). Estimates for non-cognitive skills, reported in column (4), show smaller effects. Head Start attendance increases non-cognitive skills slightly for three-year-olds (0.07 standard deviations), but the estimate for four-year-olds is negative and statistically insignificant.

In Spring 2004, members of the three-year-old cohort were still enrolled in Head Start. The cognitive point estimate for this time period is roughly comparable to the Spring 2003 estimate (0.14 standard deviations), but is less precise (s.e. = 0.08). The decline in precision between 2003 and 2004 is driven by a decline in compliance for the three-year-old cohort: many children in the control group re-applied to Head Start and were admitted at age 4, reducing the first stage from 0.68 to 0.35.[10] Similarly, the non-cognitive estimate for three-year-olds in Spring 2004 is positive, but imprecise.

The remaining rows of Table 4 show that the effects of Head Start attendance dissipate quickly once children exit the program. Cognitive estimates for both cohorts are close to zero in kindergarten (Spring 2005 for three-year-olds, Spring 2004 for four-year-olds). The estimates for four-year-olds are precise; effect sizes smaller than 0.1 standard deviations can be rejected at the 5-percent confidence level for both cognitive

[10] Head Start participation in Spring 2004 is measured from the parental survey since an administrative measure of participation is only available in Spring 2003. See the Online Appendix.

and non-cognitive skills. Estimates for both three- and four-year-olds are also statistically insignificant in 1st grade. The cognitive estimate for three-year-olds in this period is fairly large (0.10 standard deviations), but very imprecise. Non-cognitive estimates for three-year-olds are positive in Spring 2004, Spring 2005, and Spring 2006, but these estimates are not significantly different from zero. Together, these results show little evidence of cognitive or non-cognitive effects of Head Start attendance once children leave preschool.

4 Variation in Effects Across Head Start Centers

4.1 Variation in Instrumental Variables Estimates

I next turn to the primary contribution of the paper: Quantifying and explaining variation in causal effects across Head Start centers. Parametric empirical Bayes (EB) methods are the conventional approach to quantifying cross-site variation in treatment effects (Morris 1983). The EB approach involves specifying a prior distribution for the cross-site distribution of parameters, and then estimating the hyperparameters of the prior. In standard cases where site-specific estimates are unbiased and have a known sampling variance, the EB estimator takes an especially simple form: The variance of treatment effects can be consistently estimated by subtracting the average squared standard error from the sample variance of site-specific estimates (Jacob and Lefgren 2008). An efficient "shrinkage" estimator of the effect at a particular site can then be constructed as a weighted average of the estimate for that site and the overall average effect.

This simple approach is inappropriate for the HSIS data. To account for differences in compliance with random assignment across centers, it is necessary to study instrumental variables estimates rather than intent-to-treat effects of assignment to Head Start. Instrumental variables estimates have no finite moments and are not centered at the true parameter in finite samples (Nelson and Starz 1990). In addition, conventional asymptotic standard errors are likely to provide a poor approximation to their behavior in small samples (Mariano 1977). Center-specific samples in the HSIS are often small, so the finite-sample behavior of IV is relevant for center-specific IV estimates. This can be seen in Figure 1, which shows a histogram of the distribution of sample sizes across HSIS centers. More than half of centers have fewer than 10 applicants, and few have more than 25.

Table 5 illustrates the poor finite-sample behavior of center-specific IV estimates for cognitive skills in Spring 2003. The IV estimate for center j, $\hat{\beta}^j$, is the empirical analogue of equation (1). The sample standard deviation of these estimates is large (1.04 test score standard deviations), and estimates for some centers are implausible (as large as 6.8 standard deviations). Moreover, the asymptotic standard errors associated with these estimates yield nonsensical results. The average standard error is 0.9 standard deviations. As a result of large standard errors for some centers, a simple estimate of the variance of LATE formed by subtracting the average squared standard error from the sample variance of the $\hat{\beta}^j$ yields a large negative number.[11]

[11] I also used the iterative procedure for producing restricted maximum likelihood estimates of effect variation suggested by Morris (1983). This procedure places less weight on observations with large estimated sampling variances. The iterative procedure failed to converge and produced an arbitrarily large negative variance estimate.

These results make it clear that the $\hat{\beta}^j$ and their asymptotic standard errors are not informative about the extent of effect heterogeneity across centers. I next describe a framework that consistently quantifies variation in Head Start effects despite small within-center sample sizes.

4.2 Empirical Bayes Framework

My approach to quantifying effect variation uses the structure of the LATE model to describe the joint distribution of the potential outcomes $Y_i(d)$ and $D_i(z)$ conditional on Z_i and center-specific parameters. I then derive an integrated likelihood function for the sample that depends only on the hyperparameters of a parametric prior distribution, and estimate the hyperparameters. This approach circumvents the need to compute $\hat{\beta}^j$ for every Head Start center.

To describe the model, I adopt notation similar to Imbens and Rubin (1997). Let ω_g^j denote the fraction of compliance group g at center j, with $g \in \{c, nt, at\}$ indicating compliers, never takers, or always takers. Similarly, let $f_{dg}^j(y)$ denote the density function of $Y_i(d)$ for group g at center j. I assume that these potential outcome distributions belong to a common parametric family with finite-dimensional parameter vector γ, so that

$$f_{dg}^j(y) = f\left(y; \gamma_{dg}^j\right).$$

The vector of parameters at center j is therefore

$$\theta^j \equiv \left(\omega_c^j, \omega_{at}^j, \gamma_{1c}^j, \gamma_{0c}^j, \gamma_{1at}^j, \gamma_{0nt}^j\right).$$

Note that γ_{0at}^j and γ_{1nt}^j are not included in θ^j since the distribution of the data does not depend on these parameters. Similarly, ω_{at}^j is excluded since $\omega_{at}^j = 1 - \omega_{nt}^j - \omega_c^j$.

Let Y^j be the vector of observed outcomes for children at center j, and similarly for D^j and Z^j. Then the likelihood of the observed outcomes at center j conditional on the parameters at this center and Z^j is

$$\mathcal{L}\left(Y^j, D^j | Z^j; \theta^j\right) = \prod_{i:S_i=j} \left[\omega_c^j f\left(Y_i; \gamma_{1c}^j\right) + \left(1 - \omega_c^j - \omega_{nt}^j\right) f\left(Y_i; \gamma_{1at}^j\right)\right]^{D_i Z_i} \times f\left(Y_i; \gamma_{1at}^j\right)^{D_i(1-Z_i)}$$

$$\times f\left(Y_i; \gamma_{0nt}^j\right)^{(1-D_i)Z_i} \times \left[\omega_c^j f\left(Y_i; \gamma_{0c}^j\right) + \omega_{nt} f\left(Y_i; \gamma_{0nt}^j\right)\right]^{(1-D_i)(1-Z_i)}. \tag{4}$$

The cross-center distribution of parameters is described by a prior distribution G:

$$\theta^j | Z^j \sim G\left(\theta; \Omega\right).$$

The goal of empirical Bayes estimation is to estimate the hyperparameter Ω, which captures cross-center heterogeneity in outcome distributions and experimental compliance. To estimate Ω, I integrate out the site-specific parameters. The integrated likelihood is

$$\mathcal{L}^c\left(Y^j, D^j | Z^j; \Omega\right) = \int \mathcal{L}\left(Y^j, D^j | Z^j; \theta\right) dG\left(\theta; \Omega\right). \tag{5}$$

In general, the integral in equation (5) does not have a closed form. I approximate it by simulation. The simulated likelihood is given by

$$\tilde{\mathcal{L}}^c\left(Y^j, D^j | Z^j; \Omega\right) = \frac{1}{R}\sum_{r=1}^{R} \mathcal{L}\left(Y^j, D^j | Z^j; \theta_r^j(\Omega)\right),$$

where R is the number of simulation draws and $\theta_r^j(\Omega)$ is the r-th draw of the parameter vector for site j. $\theta_r^j(\Omega)$ is written as a function of Ω to emphasize that the simulation draws depend on the parameters of the prior distribution. The empirical Bayes simulated maximum likelihood estimate of Ω maximizes the sum of logarithms of simulated likelihoods across Head Start centers. This estimator is defined by

$$\hat{\Omega}_{EB} \equiv \arg\max_{\Omega} \sum_{j=1}^{J} \log \tilde{\mathcal{L}}^c\left(Y^j, D^j | Z^j; \Omega\right).$$

4.3 Empirical Bayes Estimates

Parametric Structure

To estimate the model, I assume that potential outcomes for each group are normally distributed with a common variance, so that

$$f_{dg}^j(y) = \frac{1}{\sigma}\phi\left(\frac{y-\mu_{dg}^j}{\sigma}\right).$$

The LATE at center j is therefore given by

$$\beta^j = \mu_{1c}^j - \mu_{0c}^j.$$

To capture heterogeneity in experimental compliance, let α_c^j and α_{nt}^j denote the multinomial logit utilities associated with complier and never taker probabilities at site j, defined such that

$$\omega_g^j = \frac{\exp\left(\alpha_g^j\right)}{1 + \exp\left(\alpha_c^j\right) + \exp\left(\alpha_{nt}^j\right)}$$

for $g \in \{c, nt\}$. These utilities can take any values on the real line while still ensuring the theoretical restrictions $(\omega_c^j, \omega_{nt}^j) \in [0,1]^2$ and $\omega_c^j + \omega_{nt}^j \leq 1$ are satisfied. The center-specific parameters are described by the six-dimensional vector

$$\theta^j = \left(\alpha_c^j, \alpha_{nt}^j, \mu_{1c}^j, \mu_{0c}^j, \mu_{1at}^j, \mu_{0nt}^j\right).$$

Finally, these parameters are assumed to follow a multivariate normal prior distribution:

$$\theta^j | Z^j \sim N\left(\theta_0, \Sigma\right).$$

I estimate θ_0, Σ and σ by simulated maximum likelihood, with 5,000 draws of the parameters for each Head Start center.[12]

[12]Simulated maximum likelihood estimates are asymptotically biased if the number of simulation draws grows more slowly than the square root of the sample size (Train 2003). This is unlikely to be a problem in my application since the number of draws is much larger than the number of Head Start centers, which is the relevant sample size. In practice, the estimates were stable with less than 1,000 simulation draws.

Estimates

Empirical Bayes estimates for Spring 2003 are displayed in Table 6. These estimates pool the three- and four-year-old cohorts.[13] I focus on this time period because the estimates are largest and most precise; in addition, the evidence in Chetty et al. (2011) suggests that immediate impacts of early-childhood programs may predict long-run effects better than impacts in later time periods. Results for other time periods are reported in Appendix Table A2.

The estimated prior distributions reveal significant heterogeneity in parameters across Head Start centers. Estimates of the distribution of the complier and never-taker probabilites are reported in Panel A. Consistent with the first stage estimates in Table 4, the mean compliance probability is 0.68. Compliance rates vary substantially across sites: The standard deviation of ω_c^j for the sample of childen with cognitive outcomes is 0.25. The prior distribution implies that about one-third of centers have compliance probabilities below 0.4.

Panel B of Table 6 shows estimates of the cross-center distribution of causal effects. The estimate of the average LATE for cognitive skills is 0.16 standard deviations, while the mean non-cognitive LATE is 0.062. Encouragingly, these mean impacts are similar to the pooled IV estimates in Table 4. The cross-center standard deviation of local average treatment effects, given by $\sqrt{Var\left(\mu_{1c}^j - \mu_{0c}^j\right)}$, is estimated to be 0.32 standard deviations for cognitive skills and 0.16 standard deviations for non-cognitive skills. These estimates imply substantial treatment effect variation across Head Start centers. For comparison, estimates of the standard deviations of school and teacher effectiveness are typically around 0.1 test score standard deviations (Chetty et al. 2013a; Deming 2014; Kane et al. 2008). My estimates suggest that variation in short-run Head Start effectiveness is 2 to 3 times larger than variation in value-added across teachers or schools.[14]

Figure 2 summarizes the distributions of center-specific parameters. In this figure, blue curves plot kernel densities of center-specific first stage estimates (panel A) or center-specific IV estimates (panels B and C). The densities use Epanechnikov kernels. The bandwidth for the first stage plot is 0.2, while the bandwidths for the impact plots are 0.1 standard deviations. The plots exclude IV estimates with absolute values in excess of 2 standard deviations. Red curves show the distributions of center-specific parameters implied by the empirical Bayes estimates.[15] For both the first stage and the IV estimates, estimated prior distributions show much less dispersion than distributions of center-specific estimates. Nonetheless, the estimated priors display substantively important heterogeneity.

[13] Within a center, three- and four-year-old applicants sometimes faced different probabilities of assignment to Head Start. I reweight the likelihood contributions of children within each center to account for these differences. Specifically, the likelihood contribution of child i is $\tilde{\ell}_i \equiv \ell_i^{w_i}$, where ℓ_i is the expression for the likelihood given in equation (4) and w_i is a weight proportional to child i's base HSIS weight, normalized to sum to one within each center.

[14] Raudenbush et al. (2012) discuss an alternative method for quantifying effect variation in multi-site experiments with non-compliance. Their method requires that site-specific causal effects are independent of site-specific compliance probabilities. In results available upon request, I estimated the correlation between the center-specific LATE β^j and the compliance utility parameter α_c^j. The estimate of this correlation for cognitive skills is positive and statistically significant (estimate = 0.38, s.e. = 0.15). This suggests that causal effects and compliance probabilities are not independent in the HSIS experiment.

[15] Note that this figure plots estimated prior distributions rather than distributions of center-specific posteriors. In the model used here, the posterior mean of β^j for a particular Head Start center does not have a closed form, though it is straightforward to compute this mean using Markov Chain Monte Carlo techniques. Posteriors for most centers are uninformative since the center-specific samples are very small.

To provide further context for the empirical Bayes results, I next compute the implied earnings effect of an improvement in Head Start quality, using the relationships between test score effects and lifetime earnings reported by Chetty et al. (2013b). Chetty et al. (2013b) show that a one-standard-deviation increase in teacher value-added in a single grade translates into a 1.3 percent increase in lifetime earnings. If the mapping between the short-run effect of Head Start on test scores and its effect on earnings is the same as this mapping for teachers, my results imply that a Head Start center at the 84th percentile of program quality (one standard deviation above average) will boost lifetime earnings by 3 percent relative to the average Head Start center. Assuming that children in the HSIS data will earn roughly the same amount as their parents relative to the national median (a conservative assumption since earnings revert to the mean), and using the same the assumptions on lifetime earnings trajectories described by Chetty et al. (2013b), this translates into an earnings effect of about $5,500 per child in 2010 dollars.[16] This calculation shows that the magnitude of cross-center variation in Head Start effectiveness is large enough to matter for later outcomes, and is also large relative to the per-child cost of the program (roughly $7,600; DHHS 2011).

5 Relationships Between Inputs and Effectiveness

The EB estimates reported above show that some Head Start programs are substantially more effective than others. In the remainder of the paper, I ask whether this variation in effectiveness can be explained by observed inputs. The analysis of inputs focuses on the six key variables cited by Schweinhart (2007) as responsible for the success of the Perry Preschool Project: The High/Scope curriculum, teacher education and certification, class size, instructional time, and home visting.

I investigate the relationship between inputs and causal effects using two approaches. First, I estimate interacted two-stage least squares models, with second- and first-stage equations of the form

$$Y_i = \alpha + P'_{j(i)}\phi + \beta D_i + D_i \cdot P'_{j(i)}\psi + X'_i\lambda + \epsilon_i, \tag{6}$$

$$D_i = \kappa + P'_{j(i)}\nu + \pi Z_i + Z_i \cdot P'_{j(i)}\rho + X'_i\delta + \eta_i, \tag{7}$$

where P_j is a vector of inputs and practices at center j, and $j(i)$ indexes the center to which child i applied. The first stage equations for the interactions of D_i and $P_{j(i)}$ are analogous to equation (7). This approach compares IV estimates for centers with different values of P_j . The vector ψ captures the relationship between

[16] Chetty et al. (2013b) report that the standard deviation of teacher quality is 0.13 test score standard deviations. They argue that a one-standard-deviation move upwards in this teacher quality distribution for one year raises students' earnings by 1.3 percent. The implied earnings gain per standard deviation of test scores is therefore $(1.3/0.13) = 10$ percent. I estimate that the standard deviation of Head Start quality is 0.3 test score standard deviations, so a one standard deviation increase in Head Start quality boosts earnings by $0.3 \cdot 10 = 3$ percent. Chetty et al. (2013b) estimate that the mean present value of lifetime earnings is roughly $522,000 at age 12 in 2010 dollars, which is $434,000 discounted back to age 5 at a 3-percent rate. The average HSIS family earned $17,819 per year, or 42 percent of the US median in 2002 (see http://www.census.gov/prod/2003pubs/p60-221.pdf). The average present discounted value of earnings at age 5 for children in the HSIS sample can therefore be conservatively estimated as $0.42 \cdot \$434,000 = \$182,280$. The earnings impact of a 1 standard deviation increase in Head Start quality can then be approximated as $\$182,280 \cdot 0.03 = \$5,468.40$.

the effect of Head Start attendance and observed inputs.

Second, I extend the empirical Bayes approach to incorporate dependence between inputs and causal effects in the prior distribution. Specifically, I write center-specific potential outcome means as

$$\mu_{dg}^{j} = P_j'\tau_{dg} + \xi_{dg}^{j},$$

where ξ_{dg}^{j} and the other site-specific parameters follow a joint normal distribution as in Section 4. I allow for analogous relationships between P_j and the multinomial logit utilities that generate compliance probabilities. The LATE at site j is then

$$\beta^{j} = P_j'\left(\tau_{1c} - \tau_{0c}\right) + \left(\xi_{c1}^{j} - \xi_{c0}^{j}\right). \tag{8}$$

The vector $(\tau_{1c} - \tau_{0c})$ captures the relationship between causal effects and inputs. This approach relies in part on the parametric assumptions described in Section 4, so it is likely to be less robust than two-stage least squares. The advantage of the empirical Bayes approach is that it generates an estimate of the variance of $\left(\xi_{c1}^{j} - \xi_{c0}^{j}\right)$, the residual variation in causal effects remaining after accounting for observed inputs. It can therefore be used to measure the share of effect heterogeneity explained by P_j.

Table 7 reports estimates using these two approaches. I estimate two sets of interaction models: bivariate models that include inputs in P_j one at a time, and multivariate models that include all six inputs in P_j. The estimates in columns (1) through (3) reveal that centers providing full-day service have larger cognitive effects than other centers. On average, effects of full-day centers are 0.17 standard deviations larger than effects of centers that do not offer this service. The main effect of D_i in the bivariate model for full-day service is 0.045 (s.e. = 0.047), suggesting that centers without full-day service have little effect on cognitive skills. Estimates for the multivariate interaction and maximum likelihood models are similar. This implies that the relative effectiveness of full-day centers is not explained by other inputs.

The remaining rows of Table 7 show that the other five inputs are uncorrelated with cognitive effects in Head Start. High/Scope centers do not boost scores more than non-High/Scope centers; the interaction terms associated with High/Scope are very close to zero in all models. Moreover, this difference is reasonably precisely estimated: I can reject the hypothesis that High/Scope centers are 0.15 standard deviations more effective than other centers at the 5-percent confidence level. This result weighs against the view that the High/Scope curriculum alone generated much of the success of the Perry Preschool Project (Schweinhart 2007).

Estimates of differences in cognitive effects associated with teacher education, teacher licensing, class size, and home visiting are also statistically insignificant. In the bivariate model, the interaction coefficient for the share of teachers with bachelors degrees is 0.023. This implies that a one standard deviation increase in teacher education (a 24 percentage point change) is associated with less than one hundredth of a standard deviation in increased effectiveness. This estimate is precise; the upper bound of the 95-percent confidence interval for the effect of a change of this size is 0.08 standard deviations. Similarly, the interaction coefficients associated with teacher certification are insignificant in all models, though the point estimates vary somewhat

across specifications. These results are consistent with previous evidence showing that teacher effectiveness is difficult to predict with observed characteristics (Kane et al. 2008).

The results for student/staff ratios are more surprising. Estimates from both experimental and quasi-experimental settings suggest that reduced class size boosts test scores (Chetty et al., 2011; Krueger, 1999; Angrist and Lavy 1995). In contrast, the results in Table 7 suggest that Head Start centers with larger classes are not less effective. The estimated relationship between frequency of home visiting and cognitive effects is statistically insignificant, but imprecise enough that relatively large effects cannot be ruled out.

Columns (4) through (6) of Table 7 reveal a broadly similar pattern for non-cognitive skills, with a few important differences. The High/Scope curriculum, teacher characteristics, and class size are not correlated with non-cognitive effectiveness. Unlike the cognitive estimates, however, the non-cognitive estimates for full-day service are small and insignificant. In addition, though it is not correlated with cognitive effects, frequency of home visiting seems to be related to non-cognitive effectiveness in Head Start. The two-stage least squares estimates show that centers offering more than three home visits per year boost non-cognitive skills by 0.11 standard deviations more than centers providing three or less visits. The corresponding maximum likelihood estimate is statistically insignificant, but the point estimate is positive and effects of similar size cannot be rejected. While it is not possible to unpack the mechanisms responsible for differences in cognitive and non-cognitive effects, one possibility is that cognitive effectiveness may be determined more by classroom-level factors (e.g. instructional time), while non-cognitive effectiveness may be influenced more by factors that alter the quality of the home environment (e.g. home visiting).

The final row of Table 7 reports estimates of $\sqrt{Var\left(\xi_{c1}^{j} - \xi_{c0}^{j}\right)}$, the residual standard deviation of LATE after accounting for observed inputs. Residual standard deviations are 0.295 for cognitive skills and 0.144 for non-cognitive skills. These estimates are only slightly smaller than the corresponding standard deviations in Table 6, which implies that inputs explain little of the variation in Head Start effects. More specifically, in an R^2 sense, inputs explain only 13 percent of the variation in cognitive effects, and 9 percent of the variation in non-cognitive effects.[17] Most of the heterogeneity in impacts across Head Start centers is therefore unexplained by the six inputs emphasized by Schweinhart (2007).

It is important to note that inputs and practices are not randomly assigned to Head Start centers, so the estimates in Table 7 may not reflect causal impacts of changing inputs in isolation. If the inputs analyzed here are correlated with unobserved center-level factors that influence effectiveness, these estimates will provide a misleading picture of the effects of changing inputs. Moreover, there may be important complementarities between inputs, which are not accounted for in the linear models for treatment effects implied by equations (5) and (7). Nonetheless, this analysis provides little evidence that adoption of the High/Scope curriculum or teacher credentialing requirements would improve program effectiveness in Head Start. This finding is relevant to recent policy changes that mandate increased education levels for Head Start teachers (DHHS 2008). My results show that full-day service and home visiting are most predictive of short-run Head Start

[17] The proportions of variation in cognitive and non-cognitive effects explained by inputs are $1 - \left(\frac{0.295}{0.316}\right)^2$ and $1 - \left(\frac{0.144}{0.151}\right)^2$.

effectiveness.

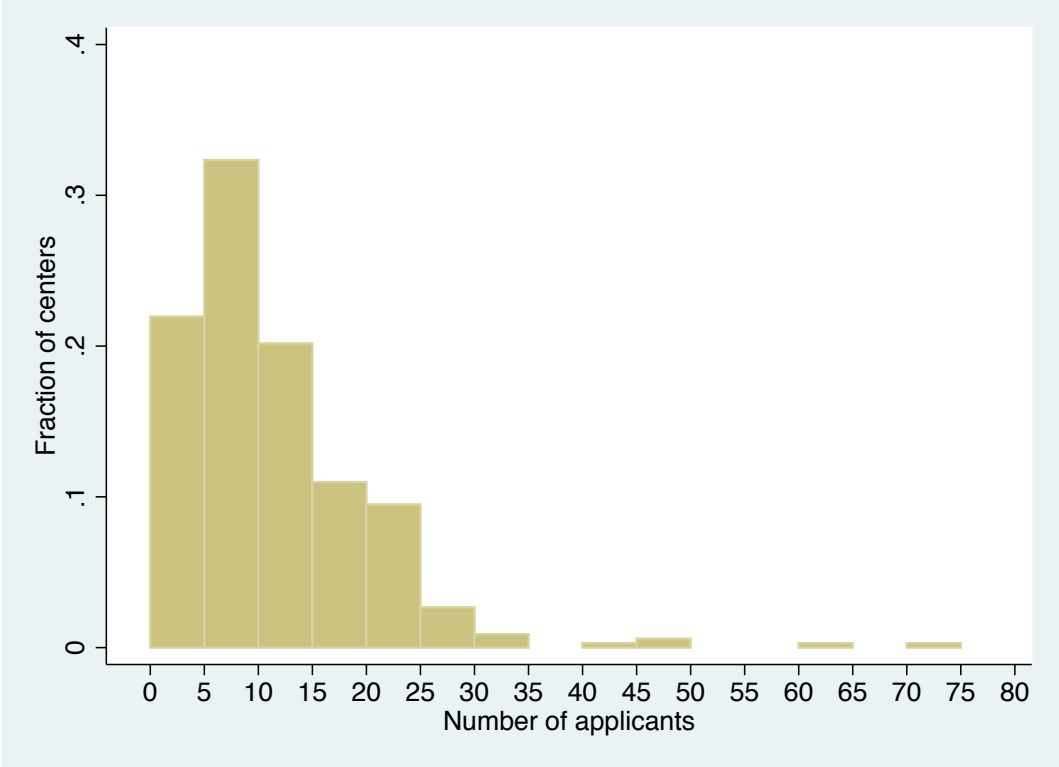
6 Conclusion

Studies of small-scale model early-childhood programs show that early intervention can dramatically boost outcomes in the short- and long-run. Randomized evidence from the Head Start Impact Study (HSIS) suggests that the Head Start program produces smaller short-run gains. This paper uses data from the HSIS to quantify impact variation across Head Start centers and ask whether differences in key inputs used by model programs can explain this variation. An empirical Bayes instrumental variables analysis reveals substantial variation in effectiveness across Head Start centers, particularly with respect to cognitive skills. Centers with full-day service and with frequent home visiting are more effective than other centers, but other inputs typically cited as important to the success of small-scale programs, including the High/Scope curriculum, teacher education and certification, and class size, do not predict program effectiveness in Head Start. Together, these inputs explain less than 15 percent of the variation in short-run cognitive effects across Head Start centers.

Much of the debate over the Head Start program centers on its long-run effectiveness (see, e.g., Gibbs et al. 2011). While the analysis here is restricted to the short run, recent evidence suggests that immediate impacts of some early-childhood programs are accurate proxies for their long-run effects (Chetty et al. 2011). Moreover, many analysts have argued that non-cognitive skills are an important channel linking short-run effects to long-term gains (Bertrand and Pan 2013; Chetty et al. 2011; Jacob 2002; Heckman et al. 2013). My estimates suggest substantial heterogeneity in short-run effects on both cognitive and non-cognitive skills, which may translate into large differences in long-run effectiveness across Head Start centers.

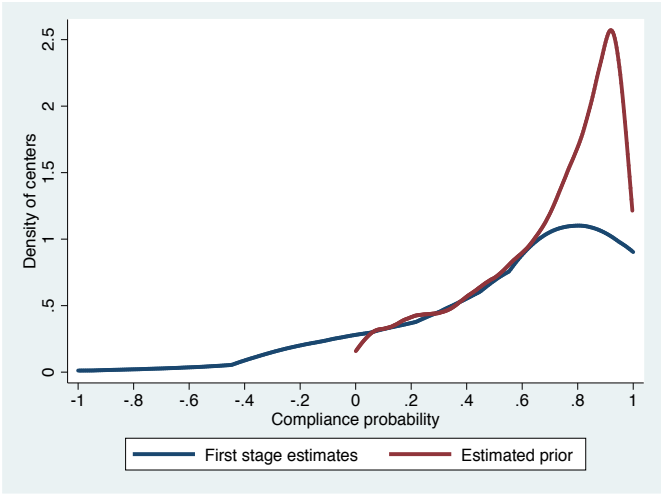
This analysis raises the further question of how the success of small-scale model programs can be replicated. This question is of immediate policy importance given recent calls to expand public preschool access (Obama 2013). The results reported here are similar to findings from the literature on teacher quality, which typically show weak relationships between observed teacher characteristics and value-added (Kane et al. 2008). These results suggest that replicating the effects of small-scale programs at a lower cost may be difficult, as the success of these programs may be due to unobserved inputs that are not easy to reproduce more cheaply and on a larger scale.

Figure 1: Histogram of Sample Sizes Across Head Start Centers

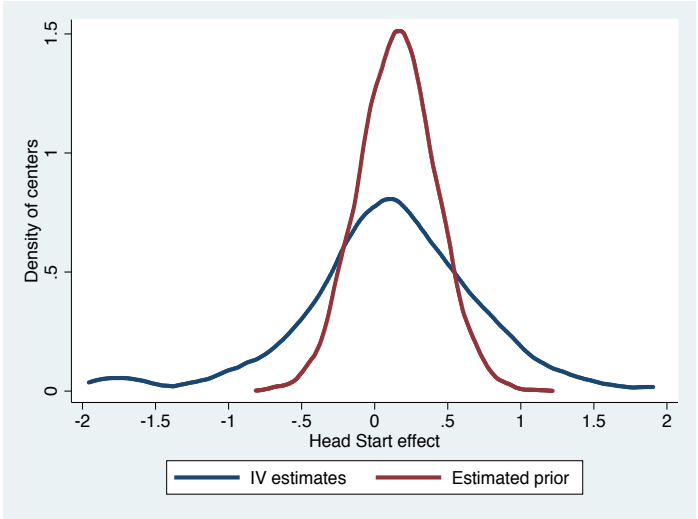


Notes: This figure shows the histogram of center-specific sample sizes in the HSIS experiment. The data are grouped into bins of 5 children (0 to 5, 6 to 10, etc.).

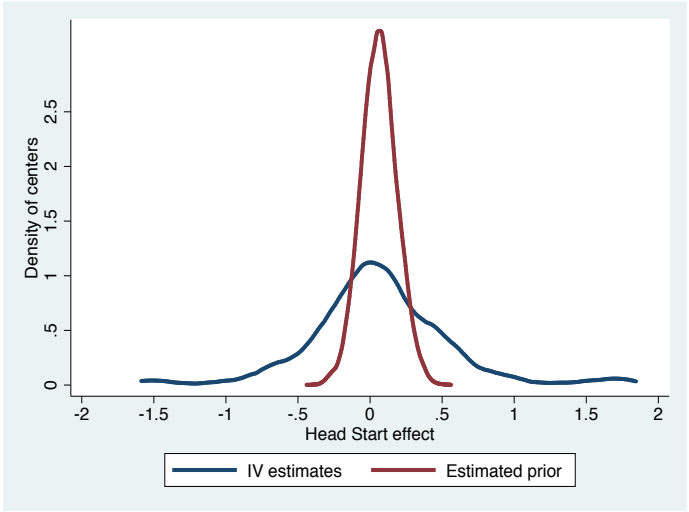
Figure 2: Distributions of Compliance Probabilities and Head Start Effects Across Centers



A. First stage



B. Cognitive impact



C. Non-cognitive impact

Notes: Blue curves show kernel density estimates of site-specific first stage and IV estimates. The impact plots exclude IV estimates greater than 2 in absolute value. Red curves show distributions of parameters implied by the empirical Bayes estimates in Table 6. These distributions are kernel densities estimated using 10,000 draws from the estimated prior distribution. All densities use an Epanechnikov kernel. The density of first stage estimates uses a bandwidth of 0.2, while the IV estimate plots use bandwidths of 0.1. The density for the first stage prior uses a bandwidth of 0.05, while the densities for the impact priors use bandwidths of 0.02.

Table 1: Outcomes Included in Summary Indices

Cognitive skills (1)	Non-cognitive skills (2)
Peabody Picture and Vocabulary Test III (PPVT adapted)	Takes care of personal things
Color names	Asks for assistance with tasks
Test de Vocabularioen Imagenes Peabody (TVIP adapted)	Helps with simple household tasks
Woodcock-Johnson III Oral Comprehension	Makes friends easily
Preschool Comprehensive Test of Phonological and Print Processing (CTOPPP)	Enjoys learning
Spanish CTOPPP	Has temper tantrums
Woodcock-Johnson III Word Attack	Cannot concentrate/pay attention for long
McCarthy Draw-a-design	Is very restless/fidgets a lot
Letter naming	Likes to try new things
Woodcock-Johnson III Letter-Word Identification	Shows imagination in work and play
Bateria R Woodcock-Munoz Identificacion de letras y palabras	Hits and fights with others
Woodcock-Johnson III Spelling	Accepts friends' ideas in playing
Bateria R Woodcock-Munoz Dictado	
Woodcock-Johnson III Applied Problems	
Woodcock-Johnson III Quantitative Concepts	
Counting Bears	
Bateria R Woodcock-Munoz Problemas Aplicados	

Notes: This table lists the cognitive and non-cognitive outcomes used in the analysis. Summary indices are averages of standardized outcomes in each category.

Table 2: Characteristics of Head Start Applicants

Variable	Control mean (1)	Differential (2)
Female	0.492	-0.011
		(0.022)
Black	0.301	0.009
		(0.018)
Hispanic	0.370	0.000
		(0.022)
Mother high school dropout	0.389	-0.022
		(0.021)
Mother some college	0.281	0.020
		(0.019)
Father high school dropout	0.353	-0.002
		(0.025)
Father some college	0.127	0.034**
		(0.015)
Teen mother	0.184	-0.023
		(0.015)
Both parents in household	0.495	-0.016
		(0.021)
Log of monthly household income	7.090	-0.037
		(0.035)
Baseline cognitive skills	-0.010	0.018
		(0.025)
Baseline non-cognitive skills	-0.016	0.016
		(0.021)
Assigned to Head Start	0.000	1.000
		-
Attended Head Start in 1st year	0.121	0.663***
		(0.017)
Joint p-value for baseline characteristics	-	0.440
N (total)	4,442	
N (completed survey)	3,577	

Notes: Column (1) shows means of baseline characteristics for Head Start applicants assigned to the control group. Column (2) shows coefficients from regressions of each characteristic on assignment to Head Start, weighting with baseline student weights. The *p*-value is from a test of the hypothesis that coefficients for all baseline characteristics are zero.

***significant at 1%; significant at 5%; *significant at 1%

Table 3: Characteristics of Head Start Centers

Variable	Mean (1)	Std. dev. (2)	Min. (3)	Max. (4)
High/Scope curriculum	0.33	0.47	0.00	1.00
Fraction of staff with bachelor's degree	0.15	0.24	0.00	1.00
Fraction of staff with teaching license	0.12	0.26	0.00	1.00
Student/staff ratio	9.81	3.99	1.54	35.00
Full day service	0.62	0.49	0.00	1.00
More than three home visits per year	0.21	0.41	0.00	1.00
Number of randomized applicants	13.70	10.66	2.00	79.00
Fraction of applicants assigned to Head Start	0.60	0.06	0.25	0.83
N (centers)		289		

Notes: This table summarizes characteristics of Head Start center in the HSIS data. The HSIS sample excludes centers where the center director did not answer the HSIS survey, and centers where the fraction of students assigned to Head Start was zero or one.

Table 4: Average Effects of Head Start on Cognitive and Non-cognitive Skills by Cohort and Year

		Cognitive skills		Non-cognitive skills	
		First stage	IV estimate	First stage	IV estimate
Time period	Cohort	(1)	(2)	(3)	(4)
Spring 2003	3-year-olds	0.679***	0.184***	0.679***	0.070**
		(0.025)	(0.039)	(0.025)	(0.031)
		2070		2062	
	4-year-olds	0.684***	0.098**	0.685***	-0.048
		(0.027)	(0.040)	(0.027)	(0.035)
		1638		1631	
Spring 2004	3-year-olds	0.351***	0.135*	0.357***	0.089
		(0.027)	(0.080)	(0.026)	(0.071)
		1980		2030	
	4-year-olds	0.694***	-0.083	0.694***	-0.037
		(0.028)	(0.052)	(0.027)	(0.040)
		1535		1555	
Spring 2005	3-year-olds	0.356***	0.049	0.360***	0.097
		(0.028)	(0.104)	(0.028)	(0.080)
		1837		1886	
	4-year-olds	0.667***	0.006	0.667***	-0.060
		(0.029)	(0.059)	(0.028)	(0.042)
		1527		1576	
Spring 2006	3-year-olds	0.348***	0.096	0.355***	0.067
		(0.028)	(0.118)	(0.027)	(0.075)
		1773		1842	

Notes: This table reports estimates of the effect of Head Start attendance on summary indices of cognitive and non-cognitive skills. Estimates come from instrumental variables models using assignment to Head Start as an instrument for Head Start attendance. These models use the HSIS baseline child weight and control for the baseline covariates from Table 2. Missing values for these covariates are set to zero, and dummies for missing values are included. Robust standard errors in parentheses.
***significant at 1%; significant at 5%; *significant at 1%

Table 5: Finite-sample Behavior of Center-Specific Instrumental Variables Estimates

	Mean	Std. dev.	Min.	Max.
	(1)	(2)	(3)	(4)
IV estimate	0.171	1.042	-4.812	6.807
IV asymptotic standard error	0.903	2.246	0.028	30.781
Implied cross-center variance of LATE		-4.77		

Notes: This table summarizes the distribution of center-specific instrumental variables estimates for cognitive skills in Spring 2003. The estimate for each center is a separate IV regression of cognitive skills on Head Start attendance instrumented by assignment, pooling the 3- and 4-year-old cohorts and using the HSIS child weights. The implied cross-center variance of LATE is the sample variance of the IV estimates minus the average squared standard error. The sample excludes centers with less than 3 applicants and centers with first stages equal to exactly zero. Two other centers with small samples and first stages very close to zero are also dropped. The sample includes 285 centers.

Table 6: Empirical Bayes Estimates

Parameter	Description	Cognitive skills (1)	Non-cognitive skills (2)
	Panel A. Compliance behavior		
$E[\omega_c^j]$	Mean compliance probability	0.689***	0.679***
		(0.018)	(0.017)
$E[\omega_n^j]$	Mean never taker probability	0.153***	0.154***
		(0.011)	(0.010)
$[Var(\omega_c^j)]^{1/2}$	Std. dev. of compliance probability	0.255***	0.224***
		(0.017)	(0.017)
	Panel B. Treatment effects		
$E[\mu_{c1}^j]$	Average treated complier mean	0.065*	0.063***
		(0.034)	(0.016)
$E[\mu_{c0}^j]$	Average control complier mean	-0.090**	0.001
		(0.040)	(0.022)
$E[\beta^j]$	Mean local average treatment effect	0.155***	0.062**
		(0.046)	(0.025)
$[Var(\mu_{c1}^j)]^{1/2}$	Std. dev. of treated complier mean	0.268***	0.097***
		(0.028)	(0.020)
$[Var(\mu_{c0}^j)]^{1/2}$	Std. dev. of control complier mean	0.351***	0.179***
		(0.045)	(0.027)
$Corr(\mu_{c1},\mu_{c0})$	Correlation between treated and control complier means	0.508***	0.624***
		(0.134)	(0.141)
$[Var(\beta^j)]^{1/2}$	Std. dev. of local average treatment effect	0.316***	0.155***
		(0.059)	(0.054)

Notes: This table shows empirical Bayes estimates of parameters of the cross-center distribution of parameters in Spring 2003. The sample pools three and four-year-old applicants. Estimates are produced via maximum simulated likelihood with 5,000 simulations for each Head Start center.

Table 7: Relationships Between Inputs and Head Start Effects

	Cognitive skills			Non-cognitive skills		
	Two-stage least squares		Maximum	Two-stage least squares		Maximum
	Bivariate	Multivariate	likelihood	Bivariate	Multivariate	likelihood
Variable	(1)	(2)	(3)	(4)	(5)	(6)
High/Scope curriculum	0.017	-0.015	0.006	0.037	0.087	0.000
	(0.064)	(0.080)	(0.078)	(0.048)	(0.055)	(0.042)
Fraction of staff with bachelor's degree	0.023	0.035	0.046	-0.026	-0.161	0.024
	(0.112)	(0.134)	(0.132)	(0.093)	(0.109)	(0.077)
Fraction of staff with teaching license	0.146	0.082	-0.121	0.143	0.096	0.022
	(0.124)	(0.149)	(0.140)	(0.095)	(0.110)	(0.079)
Student/staff ratio	0.019	0.021	-0.006	-0.008	-0.014	-0.005
	(0.016)	(0.018)	(0.007)	(0.008)	(0.009)	(0.004)
Full day service	0.169***	0.171**	0.211***	-0.005	-0.036	-0.016
	(0.060)	(0.074)	(0.068)	(0.048)	(0.054)	(0.043)
More than three home visits per year	0.039	0.020	0.098	0.114**	0.112**	0.057
	(0.064)	(0.068)	(0.085)	(0.049)	(0.051)	(0.050)
Residual std. dev. of LATE	-	-	0.295***	-	-	0.144***
			(0.062)			(0.056)

Notes: This table reports estimates of the relationship between center-specific local average treatment effects and inputs in Spring 2003. Two-stage least squares models instrument Head Start attendance and its interactions with inputs using assignment to Head Start and its interactions with inputs. Columns (1) and (4) estimate a separate single-interaction model for each input, while columns (2) and (5) include all interactions simultaneously. Main effects of interacting variables are included as controls. Two-stage least squares models weight using the HSIS baseline child weights, and also control for the baseline student characteristics from Table 2. Missing values for these covariates are set to zero, and dummies for missing values are included. Columns (3) and (6) show maximum simulated likelihood estimates from models that allow complier, always taker, and never taker means to depend on inputs. These models also allow compliance probabilities to depend on the inputs.

***significant at 1%; significant at 5%; *significant at 1%

Table A1: Attrition by Cohort and Year

		Cognitive skills		Non-cognitive skills	
		Follow-up rate	Differential	Follow-up rate	Differential
Time period	Cohort	(1)	(2)	(3)	(4)
Spring 2003	3-year-olds	0.845	0.029*	0.842	0.002
			(0.015)		(0.015)
		2449		2449	
	4-year-olds	0.822	0.030*	0.818	-0.002
			(0.018)		(0.017)
		1993		1993	
Spring 2004	3-year-olds	0.835	0.021	0.830	0.024
			(0.019)		(0.019)
		2449		2449	
	4-year-olds	0.770	0.009	0.780	-0.001
			(0.022)		(0.021)
		1993		1993	
Spring 2005	3-year-olds	0.787	-0.003	0.815	0.006
			(0.019)		(0.018)
		2449		2449	
	4-year-olds	0.766	0.010	0.791	0.012
			(0.025)		(0.024)
		1993		1993	
Spring 2006	3-year-olds	0.766	0.016	0.799	0.030
			(0.020)		(0.020)
		2449		2449	

Notes: This table reports attrition rates for the HSIS sample. Columns (1) and (3) show fractions of children with observed outcomes by cohort and time period. Columns (2) and (4) report treatment/control differences. These differences are coefficients on a treatment indicator from regressions of a dummy for an observed outcome on treatment status, with the same controls and weighting scheme as in Table 4.

***significant at 1%; significant at 5%; *significant at 1%

Table A2: Empirical Bayes Estimates by Time Period

		Cognitive skills			Non-cognitive skills		
		Spring 2003	Spring 2004	Spring 2005	Spring 2003	Spring 2004	Spring 2005
Parameter	Description	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Compliance behavior						
$E[\omega_c^j]$	Mean compliance probability	0.689***	0.497***	0.504***	0.679***	0.505***	0.439***
		(0.018)	(0.018)	(0.021)	(0.017)	(0.018)	(0.022)
$E[\omega_n^j]$	Mean never taker probability	0.153***	0.147***	0.146***	0.154***	0.143***	0.174***
		(0.011)	(0.011)	(0.010)	(0.010)	(0.011)	(0.012)
$[Var(\omega_c^j)]^{1/2}$	Std. dev. of complier probability	0.255***	0.222***	0.214***	0.224***	0.217***	0.226***
		(0.017)	(0.016)	(0.019)	(0.017)	(0.018)	(0.017)
	Panel B. Treatment effects						
$E[\mu_{c1}^j]$	Mean treated complier mean	0.065*	0.055	0.012	0.063***	0.088***	-0.278***
		(0.034)	(0.048)	(0.056)	(0.016)	(0.022)	(0.053)
$E[\mu_{c0}^j]$	Mean control complier mean	-0.090**	0.028	-0.027	0.001	0.064***	-0.162***
		(0.040)	(0.049)	(0.054)	(0.022)	(0.023)	(0.045)
$E[\beta^j]$	Mean local average treatment effect	0.155***	0.028	0.038	0.062**	0.024	-0.117**
		(0.046)	(0.064)	(0.073)	(0.025)	(0.030)	(0.055)
$[Var(\mu_{c1}^j)]^{1/2}$	Std. dev. of treated complier mean	0.268***	0.262***	0.211***	0.097***	0.079	0.110*
		(0.028)	(0.040)	(0.048)	(0.020)	(0.053)	(0.061)
$[Var(\mu_{c0}^j)]^{1/2}$	Std. dev. of control complier mean	0.351***	0.313***	0.320***	0.179***	0.168**	0.171**
		(0.045)	(0.046)	(0.049)	(0.027)	(0.066)	(0.072)
$Corr(\mu_{c1},\mu_{c0})$	Correlation between treated and control complier means	0.508***	0.662***	0.503**	0.624***	0.763***	0.546***
		(0.134)	(0.147)	(0.201)	(0.141)	(0.195)	(0.129)
$[Var(\beta^j)]^{1/2}$	Std. dev. of local average treatment effect	0.316***	0.241***	0.281***	0.155***	0.128	0.150*
		(0.059)	(0.074)	(0.095)	(0.054)	(0.079)	(0.084)

Notes: This table shows empirical Bayes instrumental variables estimates of parameters of the cross-center distribution of Head Start effects by time period. Estimates are produced via maximum simulated likelihood with 5,000 simulations for each Head Start center.

References

[1] Anderson, M. (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).

[2] Angrist, J., and Lavy, V. (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2).

[3] Berruta-Clement, J., Schweinhart, L., Barnett, W., Epstein, A., and Weikart, D. (1984). Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19. Ypsilanti, MI: High/Scope Press.

[4] Bertrand, M., and Pan, J. (2013). "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics* 5(1).

[5] Bitler, M., Domina, T., and Hoynes, H. (2012). "Experimental Evidence on Distributional Effects of Head Start." Mimeo, University of California, Irvine.

[6] Campbell, F., and Ramey, C. (1994). "Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-up Study of Children from Low-Income Families." *Child Development* 65(2).

[7] Campbell, F., and Ramey, C. (1995). "Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence: Positive Effects of Early Intervention." *American Educational Research Journal* 32(4).

[8] Cascio, E., and Schanzenbach, D. (2013). "The Impacts of Expanding Access to High-Quality Preschool Education." *Brookings Papers on Economic Activity.*

[9] Chetty, R., Hilger, N., Saez, E., Schanzenbach, D., and Yagan, D. (2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4).

[10] Chetty, R., Friedman, J., and Rockoff, J. (2013a). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-added Estimates." NBER Working Paper no. 19423.

[11] Chetty, R., Friedman, J., and Rockoff, J. (2013b). "Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood." NBER Working Paper no. 19424.

[12] Currie, J., and Thomas, D. (1995). "Does Head Start Make a Difference?" *American Economic Review* 85(3).

[13] Deming, D. (2009). "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3).

[14] Deming, D. (2013). "Using School Choice Lotteries to Test Measures of School Effectiveness." National Bureau of Economic Research Working Paper no. 19803.

[15] Epstein, A. (2007). Essentials of Active Learning in Preschool: Getting to Know the High/Scope Curriculum. Ypsilanti, MI: High/Scope Press.

[16] Fitzpatrick, M. (2008). "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement." *The B.E. Journal of Economic Analysis and Policy* 8(1).

[17] Garces, E., Thomas, D., and Currie, J. (2002). "Longer-term Effects of Head Start." *American Economic Review* 92(4).

[18] Gelber, A., and Isen, A. (2013). "Children's Schooling and Parents'Behavior: Evidence from the Head Start Impact Study." *Journal of Public Economics* 101.

[19] Gibbs, C., Ludwig, J., and Miller, D. (2011). "Does Head Start Do Any Lasting Good?" National Bureau of Economic Research Working Paper no. 17452.

[20] Gormley, W., and Gayer, T. (2005). "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program." *Journal of Human Resources* 40.

[21] Hanushek, E. (2009). "Teacher Deselection." In: Creating a New Teaching Profession, D. Goldhaber and J. Hannaway, eds. Washington, DC: Urban Institute Press.

[22] Heckman, J. (2011). "The American Family in Black and White: A Post-Racial Strategy for Improving Skills to Promote Equality." IZA Discussion Paper no. 5495.

[23] Heckman, J., Moon, S., Pinto, R., Savelyev, P., and Yavitz, A. (2010a). "The Rate of Return to the High/Scope Perry Preschool Program." *Journal of Public Economics* 94.

[24] Heckman, J., Moon, S., Pinto, R., Savelyev, P., and Yavitz, A. (2010b). "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." *Quantitative Economics* 1(1).

[25] Heckman, J., Malofeeva, L., Pinto, R., and Savelyev, P. (2013). "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103(6).

[26] Imbens, G., and Angrist, J., (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2).

[27] Imbens, G., and Rubin, D. (1997). "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25(1).

[28] Jacob, B. (2002). "Where the Boys Aren't: Non-cognitive Skills, Returns to School and the Gender Gap in Higher Education." *Economics of Education Review* 21.

[29] Jacob, B., and Lefgren, L. (2008). "Principals as Agents: Subjective Performance Assessment in Education." *Journal of Labor Economics* 26(1).

[30] Kane, T., Rockoff, J., and Staiger, D. (2008). "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6).

[31] Kling, J., Liebman, J., and Katz, L. (2007). "Experimental Analysis of Neighborhood Effects." *Econometrica* 75(1).

[32] Krueger, A. (1999). "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2).

[33] Ludwig, J., and Miller, D. (2007). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics* 122(1).

[34] Mariano, R. (1977). "Finite Sample Properties of Instrumental Variable Estimators of Structural Coefficients." *Econometrica* 45(2).

[35] Morris, C. (1983). "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78(381).

[36] Nelson, C., and Startz, R. (1990). "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58(4).

[37] Obama, B. (2013). The White House, Office of the Press Secretary. Remarks by the President in state of the union address.

[38] Raudenbush, S., Reardon, S., and Nomi, T. (2012). "Statistical Analysis for Multisite Trials Using Instrumental Variables with Random Coefficients." *Journal of Research on Educational Effectiveness* 5(3).

[39] Reynolds, A. (1998). "Extended Early Childhood Intervention and School Achievement: Age Thirteen Findings from the Chicago Longitudinal Study." *Child Development* 69(1).

[40] Schweinhart, L. (2007). "How to Take the High/Scope Perry Preschool to Scale." Paper prepared for the National Invitational Conference of the Early Childhood Research Collaborative.

[41] Schweinhart, L., Montie, J., Xiang, Z., Barnett, W., Belfield, C., and Nores, M. (2005). Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40. Ypsilanti, MI: High/Scope Press.

[42] Schweinhart, L., and Weikart, D. (1997). Lasting Differences: The High/Scope Preschool Curriculum Comparison Study Through Age 23. Ypsilanti, MI: High/Scope Press.

[43] Train, K. (2003). Discrete Choice Models with Simulation. New York: Cambridge University Press.

[44] US Department of Health and Human Services, Administration for Children and Families (2008). "Statutory Degree and Credentialing Requirements for Head Start Teaching Staff." http://eclkc.ohs.acf.hhs.gov/hslc/standards/IMs_and_PIs_in_PDF/PDF_IMs/IM2008/ACF-IM-HS-08-12.pdf. Accessed March 27, 2013.

[45] US Department of Health and Human Services, Administration for Children and Families (2010). "Head Start Impact Study, Final Report." Washington, DC.

[46] US Department of Health and Human Services, Administration for Children and Families (2011). "Head Start Program Facts, Fiscal Year 2011." http://eclkc.ohs.acf.hhs.gov/hslc/mr/factsheets/docs/hs-program-fact-sheet-2011-final.pdf. Accessed March 27, 2013.

[47] US Department of Health and Human Services, Administration for Children and Families (2012). "Third Grade Follow-up to the Head Start Impact Study." Washington, DC

[48] US Office of Head Start (2012). "Head Start Services." http://www.acf.hhs.gov/programs/ohs/about/head-start. Accessed March 27, 2013.

[49] Wong, V., Cook, T., Barnett, W., and Jung, K. (2008). "An Effectiveness-based Evaluation of Five State Pre-kindergarten Programs." *Journal of Policy Analysis and Management* 27(1).

Online Appendix

The data for this analysis come from the Head Start Impact Study (HSIS). The HSIS data includes information on 4,442 students. Each student applied to one of 353 Head Start centers in Fall 2002, and each center is associated with one of 84 regional Head Start program areas. The data includes separate files with information on test scores, answers to parental surveys, and Head Start center characteristics. This Appendix describes the procedure used to clean each data source and construct the data set used for analysis.

Test Score Data

Test score information comes from a series of assessments conducted in Fall 2002, Spring 2003, Spring 2004, Spring 2005 and Spring 2006. From each assessment file, I extract raw scores for the 17 tests listed in column (1) of Table 1. These 17 tests are the main outcomes examined by DHHS (2010). The data also include a few other tests (for example, the Leiter Sustained Attention Task), but DHHS (2010) expresses reservations about their reliability and hence they are excluded. Not all tests were administered every year, and there were some differences in the tests administered to Spanish-speaking and English-speaking students; for example, the TVIP and Spanish CTOPPP were administered to Spanish speakers only. To construct the cognitive summary index outcome, I standardize each test relative to the control group among students who took the test separately for each cohort and assessment period. I then compute the mean of observed standardized outcomes for each child. Finally, I append together the data sets for each assessment period, and use a unique student identifier to reshape the data into a wide format file with one observation per student and a separate variable for the cognitive summary index in each assessment period.

Parental Survey Data

Baseline demographics

Information on student demographics is drawn from a baseline survey of parents conducted in Fall 2002. Eighty-one percent of households responded to this survey (3,577 of 4,442). This demographic information is supplemented with a set of derived variables from the HSIS "Covariates and Subgroups" data file. This file combines the baseline survey with information collected during experimental recruitment to fill in missing values for some demographic variables. When variables are present in both files, information from the "Covariates and Subgroups" file is used.

Non-cognitive Outcomes

Indices of non-cognitive skill are constructed from the baseline parental survey and follow-up surveys conducted in Spring 2003, Spring 2004, Spring 2005 and Spring 2006. I begin with the all social and emotional outcomes analyzed by DHHS (2010). Each outcome is redefined so that a positive sign is favorable, and then standardized relative to the control group separately by cohort and survey period. I also retain raw measures of each outcome. I then append together the files for all periods. To exclude outcomes without meaningful variation, I compute the mean of each raw outcome over all survey periods, and drop outcomes

where more than 90% of responses were the same. This produces the set of outcomes listed in column (2) of Table 1. I then compute the non-cognitive summary index for each survey period as the mean of the remaining standardized outcomes. Finally, I use the unique student identifier to reshape the data into a wide format file with one observation per student and a separate variable for the non-cognitive summary index in each survey period.

Measuring Head Start Assignment and Attendance

Head Start assignment comes from an administrative variable generated at the time of random assignment. Head Start attendance in Spring 2003 is also measured administratively. To measure Head Start attendance in later periods, I combine this administrative measure with parental survey information. Specifically, I set Head Start attendance equal to one for Spring 2004, Spring 2005 and Spring 2006 if the Spring 2003 administrative measure is one, or if a parent indicated Head Start attendance at any time up to the relevant time period. For time periods after Spring 2003, the Head Start attendance variable is missing for children whose parents did not respond to the survey, because attendance cannot be accurately measured for these students. This restriction does not affect the main results, which focus on Spring 2003.

Center Characteristics

The characteristics of Head Start centers are measured from a childcare center director survey conducted in Spring 2003. The survey attempted to collect information from directors of all childcare centers attended by sample children, including members of the control group who attended childcare outside of Head Start centers in the experimental sample. The director survey data set is a student-level file, with variables capturing responses of the center director at the center attended by each child. The six key inputs listed in Table 3 are derived from the following questions:

- **High/Scope curriculum**: "If your principal curriculum has a name, what is that name?" Centers are coded as High/Scope if the director selected High/Scope from among a list of possible answers to this question.

- **Fraction of staff with bachelors degree:** "Approximately what percentage of lead and assistant teachers in your center have a bachelors degree or higher?"

- **Fraction of staff with teaching license:** "Approximately what percentage of lead and assistant teachers in your center have a teaching certificate or license?"

- **Student/staff ratio:** This variable is derived from answers to four questions. I measure center capacity using the answer to the question: "What is the center's preschool service capacity?" While capacity may not always equal enrollment, 94 percent of responses indicated that the center was filled to capacity "all the time" or "most of the time." Staff size is contructed by adding together answers to

three questions of the form: "How many X are currently employed at the center?" where X is "lead teachers," "assistant teachers," or "paid teacher aides." The student/staff ratio is then constructed by dividing capacity by staff size.

- **Full-day service:** "What child care options are provided at the center?" Centers are coded as full-day if the director selected "full-day" from a list of possible responses to this question.
- **More than three home visits per year:** "How many home visits are required per program year?" Directors were given a list of possible responses to this question. About 1 percent of responses were "1 visit," 79 percent of responses were "2-3 visits," and 20 percent of responses were "more than three visits."

I use these questions to derive the characteristics of each center of random assignment. To this end, I keep observations administratively coded as both assigned to the treatment group and attending Head Start. In some cases, codes for the center director were different for such students within a center of random assignment. I use responses for the center director most frequently associated with treated students at a given center of random assignment. For 7 percent of centers, there were two center director interviews associated with an equal number of treated students. I break ties randomly to determine which responses to use in these cases. I then keep one observation per center of random assignment. The resulting data set has information for 89 percent (314 out of 353) of centers in the HSIS experiment.

Constructing the Analysis Data Set

The procedure described above yields 5 data files: A test score file, a baseline demographic file, a non-cognitive outcome file, a file coding Head Start attendance after Spring 2003, and a center characteristics file. I merge the first four of these files using a unique student identifier. I then merge the resulting file with the center characteristics file using an identifier for center of random assignment. Finally, I merge on a sixth file containing the HSIS baseline child weights, which yields the final data set used for analysis.