Assessing the Performance of Nonexperimental Estimators for Evaluating Head Start

Andrew S. Griffen, University of Tokyo

Petra E. Todd, University of Pennsylvania

This paper uses experimental data from the Head Start Impact Study (HSIS) combined with nonexperimental data from the Early Childhood Longitudinal Study–Birth Cohort (ECLS-B) to study the performance of nonexperimental estimators for evaluating Head Start program impacts. The estimators studied include parametric cross-section and difference-in-differences regression estimators and nonparametric cross-section and difference-in-differences matching estimators. The estimators are used to generate program impacts on cognitive achievement test scores, child health measures, parenting behaviors, and parent labor market outcomes. Some of the estimators closely reproduce the experimental results, but a priori it would be difficult to know whether the estimator works well for any particular outcome. Pre-program exogeneity tests eliminate some outcomes and estimators with the worst biases, but estimators/outcomes with substantial biases pass the tests. The difference-in-differences matching estimator exhibits the best performance in terms of low bias values and capturing the pattern of statistically significant treatment effects. However, the variation in bias is greater across outcomes examined than across methods.

We thank David Card, Laura Hawkinson, Hidehiko Ichimura, Becka Maynard, Jeff Smith, seminar participants at the conference in honor of Robert LaLonde, the University of Pennsylvania Graduate School of Education, and the Tokyo Labor Economics Workshop for comments that greatly improved the paper. Andrew S. Griffen gratefully acknowledges that this research was financially supported by a
I. Introduction

There is a long-standing debate over whether the impacts of program interventions can be reliably evaluated without data from a randomized experiment. Randomization generates a control group that has the same distribution of observed and unobserved characteristics as the treatment group, allowing for straightforward attribution of causal effects. At the same time, social experimentation has drawbacks, such as high cost, the potential to distort the operation of an ongoing program, the problem of program sites refusing to participate in the experiment, the problem of randomized-out controls seeking alternative forms of treatment, and potential biases arising from nonrandom attrition. Evaluation methods that use nonexperimental data tend to be less costly and less intrusive, but a major challenge in implementing a nonexperimental evaluation strategy is choosing among the variety of estimation methods available. The choice is important given accumulated evidence that impact estimates can be highly sensitive to the identification strategy and estimator used.

This paper builds on an earlier literature that uses experimental data to evaluate the performance of nonexperimental estimators. Impact estimates based on experimental data provide a benchmark against which to gauge the performance of alternative nonexperimental estimators. Our experimental data come from the Head Start Impact Study (HSIS), and the nonexperimental data are from the Early Childhood Longitudinal Study–Birth Cohort (ECLS-B). The early literature studying the performance of nonexperimental estimators focused mainly on evaluating effects of job training programs. The subsequent literature expanded to include many other contexts, such as the effects of program interventions on health, schooling, and test score outcomes.

This paper evaluates the performance of nonexperimental estimators in assessing the effect of the Head Start early education program that is targeted at children from disadvantaged backgrounds aged 3–5. The Head Start program was previously evaluated using a large-scale social experiment that col-

---


lected the HSIS data. Our approach compares experimental impact estimates from the HSIS data to nonexperimental program impact estimates of Head Start obtained from children who attended or did not attend Head Start centers as reported in the ECLS-B data. We generate experimental and nonexperimental impact estimates on a variety of outcome measures related to children’s achievement, children’s health outcomes, parental behaviors, and parental labor supply.

In previous studies comparing nonexperimental and experimental impact estimates, it is common to generate nonexperimental estimates by combining experimental treatment group data with nonexperimental comparison group data, in part because nonexperimental treatment group data are not often available. In our case, data are available on children who participated in Head Start but were not part of the HSIS study. We generate our nonexperimental impact estimates using only the nonexperimental data for two reasons. First, if there were no experiment, a researcher would need to rely solely on nonexperimental data. As discussed in Section II, the literature on Head Start impacts includes many nonexperimental studies. Second, the reading and math cognitive test measures available in the ECLS-B data sets differ in some ways from the test measures available in the HSIS data. Carrying out the impact analysis separately for each data set maintains comparability in the way outcomes and covariates are measured for the treatment and control/comparison groups.

The plan of the paper is as follows. Section II reviews key papers on the choice among alternative nonexperimental estimators and the background literature on Head Start. Section III describes the evaluation problem and the commonly used nonexperimental estimators that we implement. Section IV describes the Head Start program and the HSIS and ECLS-B data subsamples. Section V presents the empirical results, and Section VI concludes.

II. Previous Literature

A pioneering paper in the literature that evaluates the performance of nonexperimental estimators is LaLonde (1986), which considers the evaluation of the National Supported Work (NSW) job training program. LaLonde uses data from the NSW Demonstration experiment combined with nonexperimental data from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). He applies a number of standard evaluation estimators, including simple regression adjustment, difference-in-differences, and the two-step estimator of Heckman (1979).


4. For more discussion and a re-analysis of LaLonde’s NSW data, see Dehejia and Wahba (1999, 2001) and Smith and Todd (2005a).
His findings show that alternative estimators produce a variety of estimates, most of which deviate substantially from the experimental benchmarks. Using a limited set of specification tests, LaLonde (1986) concludes that there is no good way to sort among the competing estimators and, hence, that nonexperimental methods are not generally reliable. His paper played an important role in the late-1980s movement toward using social experiments to evaluate social programs (see, e.g., Burtless 1995).

A subsequent paper by Heckman and Hotz (1989) also used the NSW data and applied a broad range of specification tests to guide the choice among nonexperimental estimators. The primary test proposed is based on pre-program data, and its validity depends on the assumption that the outcome and participation processes are similar in pre-program and post-program time periods. When their specification tests are applied to the NSW data, Heckman and Hotz (1989) find that the tests exclude the estimators that would imply substantially different qualitative conclusions (impact sign and statistical significance) than the experiment.

The more recent literature added matching estimators to the set of potential nonexperimental evaluation estimators, focusing mainly on propensity score matching methods introduced in Rosenbaum and Rubin (1983). Traditional propensity score matching pairs each program participant with a single nonparticipant, where pairs are chosen based on the degree of similarity in the estimated probabilities of participating in the program (the propensity scores). The mean impact of the program is estimated by the mean difference in the outcomes of the matched pairs. Pairwise matching methods are extended in Heckman, Ichimura, and Todd (1997, 1998) and Heckman, et al. (1998). First, kernel and local linear matching estimators are described that use multiple nonparticipants in constructing matched outcomes, reducing the asymptotic mean squared error of the estimator. Second, Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998) propose modified versions of matching estimators that can be implemented with longitudinal or repeated cross-section data. These estimators accommodate time-invariant differences between participant and nonparticipant outcomes that are not necessarily eliminated by cross-section matching.

Heckman, Ichimura, and Todd (1997, 1998) and Heckman et al. (1998) evaluate the performance of both the traditional pairwise matching estimators and cross-section and longitudinal versions of their kernel and local linear matching estimators using experimental data from the US National

---

5 Heckman and Hotz (1989) make use of somewhat different data from the NSW experiment than LaLonde (1986) does.
6 We describe the tests in Sec. V where they are applied.
7 These tests have also been applied in an evaluation context by Ashenfelter (1978), Bassi (1984), LaLonde (1986), Friedlander and Robins (1995), Raum and Torp (2002), and Regnér (2002).
Job Training Partnership Act (JTPA) Study combined with comparison group samples drawn from three sources. They show that data quality is a crucial ingredient to any reliable estimation strategy. Specifically, the estimators examined are only found to perform well when applied to data that satisfy the following criteria: (i) the same data sources (i.e., the same surveys or the same type of administrative data or both) are used for participants and nonparticipants, so that earnings and other characteristics are measured in an analogous way, (ii) participants and nonparticipants reside in the same local labor markets, and (iii) the data contain a rich set of variables relevant to modeling the program participation decision. If the comparison group data fail to satisfy these criteria, the performance of the estimators diminishes, suggesting that data quality may account for much of the poor performance of the estimators in LaLonde’s (1986) study, where participant and nonparticipant samples were located in different local labor markets and the data were collected using a combination of different survey instruments and administrative data sources.

Dehejia and Wahba (1999, 2001) use the NSW data adult male sample, one of the four subgroups originally used by LaLonde (1986), to evaluate the performance of propensity score matching methods. They find that simple matching estimators succeed in replicating the experimental NSW results and conjecture that matching-on-observables approaches exhibit lower bias than the parametric estimators that LaLonde used. However, Smith and Todd (2005a) reanalyze the NSW data and find that the better performance of matching estimators was largely due to a sample restriction that excluded a significant fraction of the observations. When these observations are included, matching estimators do not exhibit superior performance.

The early papers focused on job training programs, but the more recent debate over the performance of nonexperimental estimators also considers other types of program interventions and outcome measures related to health, school attendance, noncognitive outcomes, and test scores. The evidence on the ability of nonexperimental regression and propensity score estimators to mimic experimental impacts is mixed. Some papers find that propensity score matching performs well against an experimental comparison, particularly when there is a rich set of control variables, including baseline outcomes, and when survey instruments are similar (Diaz and Handa 2006; Handa and Maluccio 2010; Bifulco 2012). Other work argues that even when the data are of high quality and variables are measured in the same way, propensity score matching does not mimic experimental results (Agodini and Dynarski 2004; Arceneaux, Gerber, and Green 2006; Wilde and Hollister 2007; Peikes, Moreno, and Orzol 2008; McKenzie, Stillman, and Gibson 2010). Cook, Shadish, and Wong (2008) provide an overview of this literature and find that RD estimators work well in replicating experimental results (see also Black, Galdo, and Smith 2007). Consistent with the conclusions of Heckman, Ichimura, and Todd (1997, 1998) and Heckman
et al. (1998), they note that nonexperimental estimators perform better when there exists a rich set of control variables including baseline outcomes, when outcomes are measured similarly across data sets, and when the variables affecting selection into treatment are known and measured.

The approach taken in this paper is similar to that of the earlier literature in that program impact estimates based on alternative nonexperimental estimators are compared to experimental benchmarks. We consider a range of outcome measures related to child cognitive test scores, child health, parenting behaviors, and household income and labor supply. The number of outcome variables considered allows examination of how the bias varies across outcomes, across methods, and by whether the identification strategy is based on cross-sectional or longitudinal data.

Since its inception in 1965, Head Start has been a widely studied program, especially in the fields of education and psychology, with hundreds of non-experimental evaluations (McGroder 1990; Zigler and Styfco 2004; Shager et al. 2013). A fairly common finding in the earlier nonexperimental literature is that Head Start impacts on child outcomes “fade out” after the child leaves the program (Westinghouse Learning Corporation and Ohio University 1969; McKey et al. 1985; Currie and Thomas 1995, 1999). Phillips and White (2004) describe how these findings plus increasing pressure for measuring effectiveness of government programs in the 1990s led to a push for a national Head Start evaluation. The resulting study was the Head Start Impact Study (HSIS; Puma et al. 2005), which was a large-scale randomized evaluation of Head Start for children newly entering Head Start in the fall of 2002. It is the most recent and definitive study on the short-term impacts of Head Start. We use these data in our study and provide more details about the study design and data collection in Section V.

In general, the results from the HSIS showed some positive impacts at the end of the first year, especially in reading assessments and some parental behaviors, but they showed complete fade-out and statistically insignificant impacts for most outcomes by the end of first grade and continuing into third grade (Puma et al. 2012). In response to these largely disappointing findings, a series of policy reforms were implemented under the Obama administration requiring a competitive process for low-performing Head Start centers to receive funding renewal (White House 2011). The effectiveness of these policies is not clear (Gibbs et al. 2011).

One potential explanation for the modest-size experimental HSIS impact estimates is that these impacts represent the impacts of Head Start relative to
other available options, which in some cases are close substitutes for Head Start. Children in the HSIS control group have diverse child care arrangements, including care at home, with relatives, or at child care centers. Some evaluation studies use the HSIS data to estimate local average treatment effect (LATE) parameters for different subgroups and find larger gains associated with switching from the home into Head Start than from other formal child care into Head Start (Feller et al. 2014; Kline and Walters 2014). From a public finance perspective, small impacts in the cognitive domain may still pass a cost-benefit analysis because many control group children switch into Head Start from other forms of publicly financed child care (Kline and Walters 2014).

The experimental findings from the HSIS did not crowd out research that aims to estimate Head Start impacts using nonexperimental data and methods. These studies tend to use regression analysis or cross-section propensity score matching. The control variables vary by study but tend to focus on socioeconomic conditions and family and maternal characteristics, which are key determinants of Head Start eligibility. Pigott and Israel (2005) use regression estimators and Early Childhood Longitudinal Study–Kindergarten (ECLS-K) data to analyze the effects of Head Start participation on student cognitive test scores reflecting math and reading readiness upon entering kindergarten. They find that Head Start participation increases child test scores, controlling for SES and other demographics. Zhai, Brooks-Gunn, and Waldfogel (2011) use the Fragile Family and Child Wellbeing Study (FFCWS) to examine the impacts of Head Start on school readiness as measured by test scores and child behavior. They show that Head Start participation leads to higher cognitive achievement scores, higher social competence, and reduced attention problems. Using the same data, Zhai, Waldfogel, and Brooks-Gunn (2013) examine the impacts of Head Start on parenting behaviors and child maltreatment. They find that Head Start participation promotes access to learning materials in the home. Kim (2013) uses the ECLS-K and estimates the impact of Head Start on math and reading test scores at kindergarten, first grade, and third grade. The results suggest that Head Start had impacts on test scores for black students through the third grade but small or no impacts for white and Hispanic students. Lee, Brooks-Gunn, et al. (2014) use the ECLS-B and report less parental spanking for boys in Head Start but not for girls. Finally, using the same data, Lee, Zhai, et al. (2014) examine the impact of Head Start on school readiness. They find higher reading and math test scores for Head Start participants but also more conduct problems. Several of the studies (e.g., Zhai

9 There are two Early Childhood Longitudinal Studies: a study sampling the kindergarten class of 1998–99 (ECLS-K) who were followed through eighth grade and a study sampling children born in 2001 (ECLS-B) who were followed through kindergarten.

Many of these studies were published after the release of the experimental HSIS results, but these studies tend to focus on outcomes or subgroup analysis not performed in the HSIS. The continued use of nonexperimental methods even with available experimental evidence highlights the importance of understanding the performance of nonexperimental methods in evaluating Head Start. Relative to other recent research that uses the ECLS-B data, our study analyzes all the HSIS impacts that can be estimated using the ECLS-B. Our study also focuses on a quantitative assessment of the differences between experimental and nonexperimental impact estimates and examines the performance of different kinds of estimators.

There are a few studies examining the impact of Head Start on longer-term outcomes, such as educational attainment, wages, health, and crime. These papers use various approaches, including the use of sibling fixed effects (Garces et al. 2002; Deming 2009), Head Start program rules that discontinuously varied funding (Ludwig and Miller 2007), and Head Start program rules that affect eligibility (Carneiro and Ginja 2014). The studies generally find positive economically and statistically significant impacts of Head Start on longer-term outcomes. The finding of fade-out in initial cognitive impacts accompanied by significant long-term impacts is similar to results obtained with data from the Perry Preschool Project (see, e.g., Schweinhart et al. 2005; Heckman et al. 2010). Follow-up data collection for the HSIS children would provide further experimental evidence on long-term impacts.

Recent work related to Head Start examines the dynamic production of skills and investment behavior by parents (Aizer and Cunha 2012), the distribution of impacts of Head Start using the HSIS data (Bitler, Hoynes, and Domina 2014), the impact of variation in Head Start funding on outcomes (Currie and Neidell 2007), changes in parental behaviors in response to Head Start access (Gelber and Isen, 2013), out-of-sample forecasts on children’s cognitive skills and maternal labor supply of changing Head Start eligibility criteria (Griffen, 2016), and understanding what drives cross-site impacts in the HSIS (Walters 2015).

An exception, however, is Aughinbaugh (2001), which uses the National Longitudinal Survey of Youth (NLSY-97) data set and finds no long-term impacts of Head Start participation on school suspension, grade retention, and PIAT-math scores.

Some of the studies examining the impacts of Head Start on long-term outcomes come from children participating in Head Start in different time periods. Changes in Head Start funding and program practices over time raise the possibility that Head Start in different years may not have comparable impacts.
To summarize, the Head Start program has been the subject of a large number of nonexperimental investigations as well as of one large-scale randomized social experiment (HSIS). These studies examine a wide variety of short-, medium-, and long-term outcomes and analyze impacts for various subgroups. A fairly robust finding across both nonexperimental and experimental studies is that participation in Head Start yields positive benefits in the cognitive domain over the short-term (especially after 1 year). There is also fairly strong evidence that the program influences parenting behaviors. Medium-term studies, using nonexperimental data, find that cognitive test score benefits fade out in elementary school years. Nonetheless, studies of long-term impacts typically find statistically significant impacts on educational attainment, wages, health, and crime. The mechanisms through which early Head Start exposure translates into longer-term impacts are not entirely clear, but similar results have been obtained for other early childhood intervention programs (see, e.g., Heckman et al. 2010). There are no prior studies using the experimental evidence on Head Start to benchmark the performance of alternative nonexperimental estimators, which is the focus of this paper.

III. Methodology

We next describe the parametric and nonparametric evaluation estimators considered in this study.

A. The Evaluation Problem

Assessing the impact of any intervention requires making an inference about the outcomes that would have been observed for program participants had they not participated. Let $Y_1$ denote the outcome conditional on participation and $Y_0$ the outcome conditional on nonparticipation, so the impact of participating in the program is

$$\Delta = Y_1 - Y_0.$$ 

For each person, only $Y_1$ or $Y_0$ is observed, so $\Delta$ is not directly observed.

Let $D = 1$ for the group of individuals who participated in the program for whom $Y_1$ is observed. The decision to participate may involve an application and acceptance process. Let $D = 0$ for persons who do not enter the program for whom $Y_0$ is observed. Here $X$ denotes a vector of observed covariates whose distribution is assumed not to be affected by the program (e.g., gender, race/ethnicity, geographic location). The most common evaluation parameter of interest is the average impact of treatment on the treated,\[12]

---

\[12\] Following the literature, we use “treatment” and “participation” interchangeably throughout.
\[ \text{ATT}(X) = \mathbb{E}(\Delta | X, D = 1) = \mathbb{E}(Y_i - Y_0 | X, D = 1) \]
\[ = \mathbb{E}(Y_i | X, D = 1) - \mathbb{E}(Y_0 | X, D = 1), \]
\[ (1) \]
which gives the average effect of the program for program participants.\(^\text{13}\)

Experiments are designed to provide evidence on the ATT\((X)\) parameter. Data on program participants identify the mean outcome in the treated state, \(\mathbb{E}(Y_i | X, D = 1)\), and data from the randomized-out control group identify \(\mathbb{E}(Y_0 | X, D = 1)\). In nonexperimental (or observational) studies, no direct estimate of this counterfactual mean \(\mathbb{E}(Y_0 | X, D = 1)\) is available. In the next subsection, we discuss commonly used approaches for estimating the missing counterfactual mean.

**B. Nonexperimental Estimators**

Nonexperimental estimators use two types of data to impute counterfactual outcomes for program participants: data on participants prior to entering the program and data on nonparticipants. The primary estimation methods considered in this paper are parametric regression-based methods and nonparametric matching methods, both cross-sectional and longitudinal variants. We briefly review here the assumptions required to justify application of these methods, which are described in detail in Heckman et al. (1999). The cross-section estimators typically assume that program participation status is independent of outcomes conditional on a set of observed characteristics, known as a conditional independence assumption (CIA). The longitudinal estimators make an assumption that the program participation may be correlated with outcomes, even after conditioning on observables, but that the bias arising from this correlation is stable over time. This assumption is sometimes called a bias stability assumption (BSA).

**1. Cross-Section Regression Estimators**

Assume that outcome measures \(Y_{1it}\) and \(Y_{0it}\), where \(i\) denotes the individual and \(t\) the time period, can be represented by
\[ Y_{1it} = \varphi_{1i}(X_{ni}) + f_i + U_{1it}, \]
\[ Y_{0it} = \varphi_{0i}(X_{ni}) + f_i + U_{0it}, \]
\[ (2) \]
where \(f_i\) is an individual fixed effect that could represent, for example, the influences of unobserved ability or motivation on outcomes. The error terms \(U_{1it}\) and \(U_{0it}\) are distributed independently across persons and satisfy

\(^{13}\) However, there are many other parameters that may be of interest in an evaluation. See Eberwein, Ham, and LaLonde (1997), Heckman, Smith, and Clements (1997), Heckman et al. (1999), Heckman and Vytlacil (2000), Heckman (2001), and Djebbari and Smith (2008).
E(U_{it}|X_{it}) = 0 and E(U_{0i}|X_{it}) = 0. The observed outcome is \( Y_{it} = D_i Y_{1it} + (1 - D_i) Y_{0it} \), which can be written as
\[
Y_{it} = \varphi_{1i}(X_{it}) + D_i \alpha^a(X_{it}) + f_i + U_{0it},
\]
(3)
where \( \alpha^a(X_{it}) = \varphi_1(X_{it}) - \varphi_2(X_{it}) + U_{1it} - U_{0it} \) is the treatment impact. The ATT \( (X_{it}) \) parameter is
\[
\text{ATT}(X_{it}) = \varphi_{1i}(X_{it}) - \varphi_{2i}(X_{it}) + E(U_{1it} - U_{0it}|X_{it}, D_i = 1). \]
(4)
The model can be written in terms of the ATT(\( X_{it} \)) parameter by adding and subtracting \( D_i E(U_{1it} - U_{0it}|X_{it}, D_i = 1) \):
\[
Y_{it} = \varphi_{2i}(X_{it}) + D_i \text{ATT}(X_{it}) + U_{0it} \\
+ D_i [U_{1it} - U_{0it} - E(U_{1it} - U_{0it}|X_{it}, D_i = 1)].
\]
(5)
Defining the error term of the regression as
\[
\varepsilon_{it} = f_i + U_{0it} + D_i [U_{1it} - U_{0it} - E(U_{1it} - U_{0it}|X_{it}, D_i = 1)],
\]
(6)
the condition required for consistency of the cross-section regression estimator as a method for estimating ATT(\( X_{it} \)) is
\[
E(\varepsilon_{it}|X_{it}, D_i = 1) = 0,
\]
(7)
which requires that
\[
E(f_i + U_{0it}|X_{it}, D_i = 1) = 0. \]
(8)
This assumption implies that individuals are not selecting into the program on the basis of unobserved characteristics \( f_i \) or on the basis of \( U_{0it} \). This assumption is essentially a CIA, expressed within the context of an econometric model for outcomes.

2. Difference-in-Differences Regression Estimators

A difference-in-differences (DID) estimator measures the impact of the program by the post-program and pre-program difference in outcomes between participant. It uses both pre-program and post-program data (\( t \) and \( t' \) data) on \( D = 1 \) and \( D = 0 \) persons. The outcomes in the post-program and the pre-program time periods can be written as
\[
Y_{it} = \varphi_{2i}(X_{it}) + D_i \text{ATT}(X_{it}) + f_i + U_{0it} \\
+ D_i(U_{1it} - U_{0it} - E(U_{1it} - U_{0it}|D_i = 1, X_{it})).
\]
(9)
\[
Y_{it'} = \varphi_{2i'}(X_{it'}) + f_i + U_{0it'}.
\]
(10)
Taking differences, we get

\[ Y_{it} - Y_{i't} = \varphi_{it}(X_{it}) - \varphi_{i't}(X_{i't}) + D_{it}\text{ATT}(X_{it}) + (U_{0it} - U_{0i't}) + \]

\[ D_i(U_{1it} - U_{0it}) - E(U_{1it} - U_{0it} | D_i = 1, X_{it}). \]  

(11)

(12)

Defining the error term as

\[ \hat{\varepsilon} = (U_{0it} - U_{0i't}) + D_i(U_{1it} - U_{0it} - E(U_{1it} - U_{0it} | D_i = 1, X_{it})), \]

the estimator is consistent if

\[ E(\hat{\varepsilon} | D_i = 1, X_{it}) = 0, \]

which requires that

\[ E(U_{0it} | D_i = 1, X_{it}) = E(U_{0i't} | D_i = 1, X_{it}). \]

This assumption is sometimes called the bias stability assumption.

LaLonde (1986) implements both the standard DID estimator just described and a version that includes \( Y_{i't} \) as a right-hand-side variable, relaxing the implicit restriction in the standard DID estimator that the coefficient associated with lagged \( Y_{i't} \) equals 1.\(^{14}\) We present estimates from both the standard DID and the lagged value-added estimators.

C. Cross-Section Matching Methods

Traditional matching estimators pair each program participant with an observably similar nonparticipant and interpret the difference in their outcomes as the effect of the program (see, e.g., Rosenbaum and Rubin 1983). To justify their application, researchers often assume CIA, namely, that there exists a set of observable conditioning variables \( Z \) (which may be a subset or a superset of \( X \)) for which the nonparticipation outcome \( Y_0 \) is independent of participation status \( D \) conditional on \( Z \).

\[ Y_0 \perp D | Z. \]  

(13)

It is also assumed that for all \( Z \) there is a positive probability of either participating \((D = 1)\) or not participating \((D = 0)\); that is,

\[ 0 < \Pr(D = 1 | Z) < 1. \]  

(14)

\(^{14}\) LaLonde (1986) called the model with the lagged outcome as a right-hand side variable an “unrestricted DID estimator.”

\(^{15}\) In the terminology of Rosenbaum and Rubin (1983), treatment assignment is “strictly ignorable” given \( Z \). To simplify notation, in this section, we suppress the individual and time subscripts.
This assumption implies that a match on $Z$ can be found for all $D = 1$ persons. If assumptions (13) and (14) are satisfied, then, after conditioning on $Z$, the $Y_0$ distribution observed for the matched nonparticipant group can be substituted for the missing $Y_0$ distribution for participants.

As noted by Heckman, Ichimura, and Todd (1997, 1998) and Heckman et al. (1998), assumption (13) is overly strong if the parameter of interest is the average impact of treatment on the treated (ATT($X$)), in which case conditional mean independence suffices:

$$E(Y_0|Z, D = 1) = E(Y_0|Z, D = 0) = E(Y_0|Z).$$

Furthermore, to estimate the ATT($X$) parameter, we require only the possibility of a nonparticipant analogue for each participant, so the required condition is

$$\Pr(D = 1|Z) < 1.$$  

Under these assumptions—either (13) and (14) or (15) and (16)—the mean impact of the program can be written as

$$\Delta = E(Y_1 - Y_0|D = 1),$$

$$= E(Y_1|D = 1) - E_{Z,D=1}\{E_Y(Y_0|D = 1, Z)\},$$

$$= E(Y_1|D = 1) - E_{Z,D=1}\{E_Y(Y_0|D = 0, Z)\},$$

where the first term can be estimated from the treatment group and the second term from the mean outcomes of the matched (on $Z$) comparison group.

In a social experiment, (13) and (14) are satisfied by virtue of random assignment of treatment. For nonexperimental data, there may or may not exist a set of observed conditioning variables for which the conditions hold. A finding of Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998) in their application of matching methods to the JTPA data and of Dehejia and Wahba (1999, 2001) in their application to the NSW data is that (14) was not satisfied, meaning that for a fraction of program participants, no match could be found. If there are regions where the support of $Z$ does not overlap for the $D = 1$ and $D = 0$ groups, then matching is only justified when performed over the common support region. The estimated treatment effect must then be redefined as the treatment impact for program participants whose $p$-scores lie within the overlapping support region.

16 An advantage of experiments noted by Heckman (1997), as well as by Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998), is that they guarantee that the supports are equal across treatments and controls, so the mean impact of the program can be estimated over the entire support.
1. Reducing the Dimensionality of the Conditioning Problem

Matching may be difficult to implement when the set of matching variables Z is large.\textsuperscript{17} Rosenbaum and Rubin (1983) prove a result that is useful in reducing the dimension of the conditioning problem. They show that for random variables Y and Z and a discrete random variable D,

\[ E(D|Y, \Pr(D = 1|Z)) = E(E(D|Y, Z)|Y, \Pr(D = 1|Z)), \]

so \( E(D|Y, Z) = E(D|Z) = \Pr(D = 1|Z) \) implies that \( E(D|Y, \Pr(D = 1|Z)) = E(D|\Pr(D = 1|Z)) \). This implies that when \( Y_0 \) outcomes are independent of program participation conditional on Z, they are also independent of participation conditional on the propensity score, \( \Pr(D = 1|Z) \). Provided that the propensity score can be estimated parametrically (or semi-parametrically at a rate faster than the nonparametric rate), the dimensionality of the matching problem is reduced by matching on the univariate propensity score. For this reason, much of the recent evaluation literature on matching focuses on propensity score matching methods.\textsuperscript{18}

2. Matching Estimators

For notational simplicity, let \( P = \Pr(D = 1|Z) \). A typical matching estimator takes the form

\[ \hat{\alpha}_M = \frac{1}{n_1} \sum_{i \in I_1 \cap S_p} [Y_{1i} - \hat{E}(Y_{0i}|D = 1, P_i)], \]  

(17)

where

\[ \hat{E}(Y_{0i}|D = 1, P_i) = \sum_{j \in I_0} W(i, j) Y_{0j}, \]

and where \( I_1 \) denotes the set of program participants, \( I_0 \) the set of nonparticipants, \( S_p \) the region of common support (see below for ways of constructing this set), and \( n_1 \) the number of persons in the set \( I_1 \cap S_p \). The match for each participant \( i \in I_1 \cap S_p \) is constructed as a weighted average over the outcomes of nonparticipants, where the weights \( W(i, j) \) depend on the distance between \( P_i \) and \( P_j \). Alternative matching estimators differ in how the weights \( W(i, j) \) are constructed. In this paper, we implement

\textsuperscript{17} If Z is discrete, small cell problems may arise. If Z is continuous and the conditional mean \( E(Y_i|D = 0, Z) \) is estimated nonparametrically, then convergence rates will be slow due to the “curse of dimensionality” problem.

\textsuperscript{18} Hahn (1998), Heckman, Ichimura, and Todd (1998), and Angrist and Hahn (2001) consider whether it is better in terms of efficiency to match on \( P(X) \) or on \( X \) directly. For the TT parameter, neither is necessarily more efficient than the other. If the treatment effect is constant, then it is more efficient to condition on the propensity score.
nonparametric matching estimators as developed in Heckman, Ichimura, and Todd (1998).

**Kernel and Local Linear Matching.**—Nonparametric matching estimators construct a match for each program participant using a weighted average over multiple persons in the comparison group. The weighting function typically gives more weight to nonparticipant observations that have propensity scores that are close to that of each participant, $P_i$, and less weight (or zero weight) to observations that are less close. Consider, for example, the kernel matching estimator described in Heckman, Ichimura, and Todd (1997, 1998) and Heckman et al. (1998), given by

$$
\hat{\alpha}_{KM} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_{1i} - \frac{\sum_{j \in I_0} Y_{0j} G \left( \frac{P_j - P_i}{a_n} \right)}{\sum_{j \in I_0} G \left( \frac{P_j - P_i}{a_n} \right)} \right\},
$$

where $G(\cdot)$ is a kernel function and $a_n$ is a bandwidth parameter. In terms of equation (17), the weighting function, $W(i, j)$, is equal to

$$
\frac{G \left( \frac{P_j - P_i}{a_n} \right)}{\sum_{k \in I_0} G \left( \frac{P_k - P_i}{a_n} \right)}.
$$

Under standard conditions on the bandwidth and kernel,

$$
\frac{\sum_{j \in I_0} Y_{0j} G \left( \frac{P_j - P_i}{a_n} \right)}{\sum_{k \in I_0} G \left( \frac{P_k - P_i}{a_n} \right)}
$$

is a consistent estimator of $E(Y_0|D = 1, P_i)$.\(^{19}\)

In this paper, we implement a generalized version of kernel matching, called local linear matching. The local linear weighting function is given by

$$
W(i, j) = \frac{G_{ij} \sum_{k \in I_0} G_{ik} (P_k - P_i)^2 - \left[ G_{ij} (P_j - P_i) \right] \left[ \sum_{k \in I_0} G_{ik} (P_k - P_i) \right]}{\sum_{j \in I_0} G_{ij} \sum_{k \in I_0} G_{ik} (P_k - P_i)^2 - \left( \sum_{k \in I_0} G_{ik} (P_k - P_i) \right)^2}. \quad (18)
$$

As shown in Fan (1992a, 1992b), the expression for the standard kernel regression estimator can be obtained from a kernel weighted regression of $Y_{0j}$ on a constant term. The local linear estimator can be obtained from a weighted regression of $Y_{0j}$ on a constant term and on $(P_j - P_i)$. Research by Fan (1992a, 1992b) demonstrated advantages of local linear estimation over more standard kernel estimation methods. These advantages include

\(^{19}\) We assume that $G(\cdot)$ integrates to one and has mean zero and that $a_n \to 0$ as $n \to \infty$ and $n a_n \to \infty$. In estimation, we use the quartic kernel function, $G(s) = (15/16)(s^4 - 1)^2$ for $|s| \leq 1$, else $G(s) = 0$. 
a faster rate of convergence near boundary points and greater robustness to
different data design densities. 20

Determining the Support Region.—To implement the matching estimator
given by equation (17), the region of common support $S_p$ needs to be deter-
dined. By de
finition, the region of support includes only those values of $P$
that have positive density within both the $D = 1$ and $D = 0$ distributions.
The common support region can be estimated by

$$\hat{S}_p = \{ P : \hat{f}(P|D = 1) > 0 \text{ and } \hat{f}(P|D = 0) > 0 \},$$

where $\hat{f}(P|D = d), d \in \{0, 1\}$, are nonparametric density estimators given
by: 21

$$\hat{f}(P|D = d) = \sum_{k \in I_d} G\left( \frac{P_k - P}{\delta_n} \right).$$

To ensure that the densities are strictly greater than zero, we require that the
densities be strictly positive and exceed zero by a threshold amount deter-
dined by a “trimming level” $q$. After excluding any $P$ points for which the
estimated density is exactly zero, we exclude an additional $q\%$ of the re-
main ing $P$ points for which the estimated density is positive but very low.
The set of eligible matches are given by

$$\hat{S}_p = \{ P \in I_1 \cap \hat{S}_p : \hat{f}(P|D = 1) > c_q \text{ and } \hat{f}(P|D = 0) > c_q \},$$

where $c_q$ is the density cut-off trimming level. 22

Difference-in-Differences Matching.—As previously noted, cross-section
matching estimators assume CIA, which is a strong assumption on the pro-
gram participation process. For a variety of reasons there may be systematic
differences between participant and nonparticipant outcomes, even after
conditioning on observed characteristics. Such differences may arise, for
example, (i) because of program selectivity on unmeasured characteristics,

20 Specifically, these nonparametric estimators have a finite-sample bias term that
goes to zero asymptotically. The expression for the bias term for the local linear
estimator does not depend on the density of the data ($f(P)$), which is considered
to be a source of robustness. The variance of the kernel and the local linear estima-
tors are identical. See Fan (1992a, 1992b) and Fan and Gijbels (1996).
21 In implementation, we select a fixed, global bandwidth parameter using cross-
validation, as described in Frölich (2004) and Galdo, Smith, and Black (2008).
22 The $q$th quantile $c_q$ is determined by solving for

$$\sup_{c_q} \frac{1}{2J} \sum_{\{ad/\hat{S}_p\}} \left\{ 1 \left( \hat{f}(P|D = 1) < c_q \right) + 1 \left( \hat{f}(P|D = 0) < c_q \right) \right\} \leq q,$$

where $J$ is the number of observed values of $P$ that lie in $I_1 \cap \hat{S}_p$. That is, matches
are constructed only for the program participants for which the propensity scores
lie in $\hat{S}_q$. 

and (ii) because of levels differences in outcomes participants and nonparticipants, arising perhaps because of differences in how the outcomes are measured (as when data are collected using different survey instruments).

A difference-in-differences (DID) matching strategy, as described in Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998), allows for temporally invariant differences in outcomes between participants and nonparticipants.

The DID propensity score matching estimator requires a bias stability assumption (BSA), namely, that

$$E(Y_{0t} - Y_{00}|P, D = 1) = E(Y_{0t} - Y_{00}|P, D = 0),$$

where \(t\) and \(t'\) are time periods after and before the program enrollment date. This assumption is analogous to that required by the standard DID regression estimator, but the DID matching estimator does not impose a parametric model on the outcome variable.

The DID matching estimator also requires the support condition given in (16), which must hold in both periods \(t\) and \(t'\). The nonparametric difference-in-difference estimator is given by

$$\hat{\alpha}_{DDM} = \frac{1}{n_1} \sum_{i \in I \cap S_y} \left\{ (Y_{1i} - Y_{0i}) - \sum_{j \in I \cap S_y} W(i, j)(Y_{0j} - Y_{0j}) \right\},$$

where the weights can correspond to either the kernel or the local linear weights defined above. An important difference between the DID matching estimator and the standard DID regression estimator is that the matching estimator reweights the nonparticipant observations according to the proximity of their propensity scores to those of treatment group observations.

IV. Data

As previously described, we evaluate the performance of alternative nonexperimental estimators against experimental benchmarks. The experimental data come from the Head Start Impact Study (HSIS), which was a randomized control trial of Head Start. The nonexperimental comparison data are subsamples from the Early Childhood Longitudinal Study–Birth Cohort (ECLS-B). We next describe the HSIS experiment, the two data sets we use, and the outcome variables we analyze.

A. Head Start Impact Study

Head Start is a preschool program funded by the federal government that serves nearly 1 million children between the ages of 3 and 5 from low-income families in the United States and has an explicit goal to “promote school readiness by enhancing the social and cognitive development of children.” Despite long-term popular and political support for Head Start since its inception in 1965, the 105th US Congress in 1998 took the
unusual step to mandate a randomized evaluation of Head Start (P.L. 105–285). This law led to the development and implementation of the HSIS (Puma et al. 2005), a randomized evaluation of the Head Start among new applicants in the fall of 2002. Head Start centers were selected for inclusion in the experiment according to a stratified randomized sampling design that was intended to be largely representative of the entire population of Head Start centers. However, to randomize the offer of Head Start but still preserve the number of enrollees, the centers were required to be oversubscribed. Although 85% of the centers satisfied this criteria, this restriction potentially compromised the representativeness of the sample. In total there were 2,783 children randomized to receive an offer of Head Start and 1,884 children randomized out to serve as a control group.

The HSIS actually consisted of two interventions depending on the child’s age: a group of 4-year-olds who were randomized to receive an offer of Head Start or not (HSIS 4-year-olds) and a group of 3-year-olds who were randomized to receive an offer of Head Start or to receive a delayed option to reapply for Head Start the subsequent year (HSIS 3-year-olds). The study followed children longitudinally and collected follow-up data at the end of the first Head Start year, at the end of the second Head Start year (for the HSIS 3-year-olds), and at the end of kindergarten, first grade, and third grade. We focus on the first-year impacts in the current study. This gives us two groups receiving the offer of Head Start for 1 year.

The HSIS reported first-year impact estimates of Head Start on 33 outcomes measures across four domains: cognitive, socioemotional, health, and parenting. As described below, we can only find comparable outcomes in the ECLS-B data for a subset of the measures. Outcomes in the cognitive and socioemotional domains fall clearly within Head Start’s stated mission. The justification for examining impact estimates in the health and parenting domains comes from Head Start’s outreach to parents, including home visits, providing information to parents about child development and parenting, and connecting parents to local services for children. Although parental labor supply and household income were not originally outcomes considered by the HSIS, Head Start creates some potential labor market incentives from its role as a child care service provider and some labor supply disincentives stemming from income eligibility cutoffs (Griffen 2016).

The HSIS found impacts in some domains that were statistically significantly different from zero, but, as previously noted, these impacts tended to fade out over time. Of the 33 outcomes considered in the first-year findings, nine impact estimates were statistically significant for the 3-year-olds and seven impact estimates were statistically significant for the 4-year-olds. For the first-year impacts, all of the statistically significant impacts estimates had the “right” sign, showing improvement in the expected direction for the treatment group relative to the control group. Outcomes with statistically significant impacts included early reading assessments in the cognitive do-
main, some health outcomes, and some parenting behaviors. The impact estimates were relatively large in magnitude. Relative to between-group gaps observed in the ECLS-B baseline data, impact estimates ranged between 28% to 50% of the black-white gap and 21% to 71% of the gap between mothers with 12 versus 16 years of education. However, impact estimates measured at a later time—at kindergarten entry, the end of first grade, and the end of third grade—were mostly statistically insignificant, and the few impacts that were found did not show a clear pattern of favorable or unfavorable impacts (Puma et al. 2012).

B. Early Childhood Longitudinal Study–Birth Cohort

For our nonexperimental analysis, we focus on subsamples from the ECLS-B, a nationally representative panel of 14,000 children born in 2001. Researchers followed the children from birth until kindergarten entry and collected detailed information about their family background, parental labor supply, household income, parental behaviors, child care usage, and cognitive achievement outcomes. The data have several unique features that make them a good comparison data set for the HSIS. First, the ECLS-B covers the 2001 cohort of children born in the United States as compared to children in the HSIS who were born from 1997 to 1999. The use of nearby age cohorts allows us to compare children exposed to Head Start under similar program conditions. Second, the ECLS-B drew some sample questions directly from the HSIS, which increases the likelihood of finding similar questions across data sets. Third, the ECLS-B collected extensive longitudinal data on child outcomes, home conditions, and child care participation, which provides a rich data source to control for selection into Head Start and for prior achievement (Hawkinson et al. 2013). An additional useful feature of the ECLS-B is that the study sent questionnaires directly to child care providers, so a program’s status as a Head Start program can be verified. Provider reports are important because, as Kim (2013) argues using the ECLS-K, parental reports of Head Start overstate participation relative to national statistics, and this reporting error appears to be nonrandom. We find similar patterns in the ECLS-B.

The ECLS-B consists of five rounds of data collection. The researchers visited the children when they were approximately 9 months, 2 years, 4 years, and 5 years old, with an additional follow-up round for delayed kindergarten entrants. In addition to the irregular spacing between rounds, there is also variability in age of assessment at each round. For example, in the third round, the children ranged in age from 44 months to 65 months. There are several issues when constructing comparison groups for the HSIS 3-year-olds and 4-year-olds. First, using the child care provider interview in the third round, we observed whether a child in the ECLS-B was participating in Head Start at a particular moment in time. In addition to the variation in age at the round that may place children into different cohorts, there is also...
variation in how long children had been in Head Start. That is, when we observed them participating in Head Start, some of the children had already been participating in Head Start for some time, while others would have just recently begun.

To address this issue, we combine the information from the provider Head Start reports with information from the parental survey about when the child first began to participate in Head Start. To mimic the design of the HSIS baseline data collection, we assign children to the 4-year-old Head Start group (ECLS-B 4-year-olds) if they both reported being in Head Start and had begun Head Start within 2 months of the ECLS-B third round. Other children who began Head Start earlier are assigned to the 3-year-old cohort (ECLS-B 3-year-olds). For the ECLS-B 3-year-olds, we use the third round for their endline outcomes, and for the ECLS-B 4-year-olds we use the fourth round for their endline outcomes.

We estimate our propensity score (jointly for the 3-year-old and 4-year-old children) as the probability of being in Head Start in the third round of the ECLS-B. Children are eligible for Head Start if they are between the ages of 3 and 5 and if their family’s income is equal to or below the poverty line. These poverty line cutoffs are not strict, however, as Head Start centers can enroll children above the cutoffs depending on their assessment of the child’s need. The ECLS-B data include household income, the main determinant of eligibility, and an extensive set of other covariates measured at baseline.

C. Outcome Measures

Despite the involvement of similar stakeholders and the purposeful reuse of some measures across the HSIS and the ECLS-B, outcome measures are similar in only some of the domains. In the cognitive domain, the HSIS reports 14 different outcome measures, whereas the ECLS-B simply reports reading and math scores. Of these 14 cognitive domain assessments in the HSIS, six can be converted into percentile scores for a nationally representative sample using conversion factors provided in the HSIS data. Using the converted percentile scores, we simply average the five percentile scores

---

23 Children can also be eligible if their family is eligible for public assistance or if the child is homeless or in foster care. The Head Start website has a description of these income eligibility rules. The online appendix for Carneiro and Ginja (2014) has an extensive discussion of the eligibility criteria.

24 The income used to determine eligibility for Head Start is the income from the parent(s) or the legal guardian(s) of the child. Any other household income does not count toward eligibility. In the ECLS-B, more than 90% of households have only parents or legal guardians working and supporting the child.

25 These scores consist of response items drawn from the PPVT, the Pre-CTOPPP, the PreLAS 2000, and the TEMA-3.

26 These are the PPVT-Adapted, WJ III Letter-Word Identification, WJ III Spelling, WJ III Oral Comprehension, and WJ III Pre-Academic Skills under the category of reading and WJ III Applied Problems under the category of math.
in the reading area to form our HSIS reading percentile score, and we use the WJ III Applied Problems for our math percentile score. In the ECLS-B, the exact combination of items for the math and reading assessments is not given, but we do know that some of the assessment items were borrowed from items used in the HSIS. Because the ECLS-B is nationally representative, we directly convert these reading and math scores into national percentile scores.

In the socioemotional domain, HSIS reported impact estimates for nine outcomes, but the ECLS-B has no comparable questions. In the health domain, two out of five HSIS outcomes were measured with similar questions in ECLS-B. Specifically, both data sets have information on parental reports about whether the child’s overall health is good/excellent and whether the child has health insurance coverage.27

In the parenting domain, four out of the five outcomes were similar: whether the parent read to the child in the last week, a parental safety practices scale consisting of car seat and smoke detector use, whether the parent spanked the child in the last week, and whether the parent used a time out in the last week.28 The wording of the questions sometimes differs slightly across the studies.29

Finally, in the labor domain, both data sets include measures of maternal employment, paternal employment, and household income. We define maternal and paternal employment as equal to one if the respondent reported working full-time or part-time and zero otherwise.30 For household income, the HSIS asks for the income of everyone in the house for the last month, whereas the ECLS-B asks for income of everyone in the house over the past year.31 In comparing impact estimates across experimental and nonexperimental estimators, the child health question is nearly identical across the data sets. The health insurance question for the ECLS-B has more prompts for different kinds of insurance.27 The HSIS used a parental safety scale with an index of four different measures, but the ECLS-B had similar questions for only two of the scale’s four measures. We take the two comparable questions from the ECLS-B and the HSIS, and we construct a modified parental safety practices scale using the same aggregation as in the original HSIS variable construction.28

For example, for car seat use in the safety practices scale, the HSIS asks about using a safety seat or a seat belt, whereas the ECLS-B asks about only a car seat. Questions about spanking in the ECLS-B include an initial prompt stating that “sometimes kids mind pretty well and sometimes they don’t,” and some questions have different scales that need to be converted to be comparable (e.g., the HSIS asks whether the parent used time out, and the ECLS-B asks about the number of times the parent used time out).29

For household income, the HSIS contains an additional prompt for being in the military, which we also classify as employment.30 The surveys also differ in that the HSIS first asks for the amount directly before prompting for a bracketed response (in case of nonresponse or refusal), whereas the ECLS-B asks the respondent to locate within a bracket and then prompts for more detailed income information for poorer individuals.
mental estimators, differences in survey instruments could be a potential source of bias.

V. Empirical Results

Table 1 shows descriptive statistics for children age 3 and age 4 in the HSIS and in the ECLS-B subsamples. In the ECLS-B, we exclude children from households with income above $50,000 to approximate the income eligibility rules for Head Start. We provide more details below when discussing the propensity score estimation.32 For the ECLS-B sample, we show separate statistics for Head Start participants and nonparticipants. The percentage of children who are black is lower among 3-year-olds in the HSIS sample compared to the ECLS-B sample (29% vs. 35%) but higher among the 4-year-olds (21% vs. 16%). The HSIS subsample has a lower percentage of Hispanics—28% of 3-year-olds in HSIS compared to 32% in ECLS and 34% of 4-year-olds compared to 45%. Children in the HSIS sample are about 3 months younger on average than in the ECLS-B sample (at the time of observation). Maternal education levels are slightly lower in the HSIS sample. Relative to the Head Start participants, the ECLS-B nonparticipants have fewer minorities, more married families, higher maternal education, and higher family income.

Table 2 shows the mean outcome measures at baseline for the HSIS sample and for the ECLS-B Head Start participant subsamples. Cognitive test scores are only available at baseline for the 4-year-old sample. Reading and math percentile scores are remarkably similar. The average reading and math percentile scores are 31 and 33 for the HSIS compared to 32 and 34 for ECLS-B. Children in the ECLS-B Head Start and HSIS participant sample are similar in terms of the likelihood of the child being in excellent or good health according to parent reports, but the children in the ECLS-B are more likely to be covered by health insurance. Table 2 also compares parenting behaviors. The rate of reading to the child is lower among ECLS-B Head Start participant children than among HSIS children. Parents in the HSIS sample and the ECLS-B Head Start participant sample are comparable in parental safety practices (car seat and smoke detector use). Among 3-year-olds, the use of time out is higher in the ECLS-B subsample, but the use of spanking is higher in the HSIS sample. The reverse holds in the 4-year-old samples. Household income is substantially lower in the HSIS sample. For 3-year-olds, it is $17,286 in HSIS compared to $22,375 for the ECLS-B Head Start participants. For 4-year-olds, it is $18,672 in HSIS and $25,573 in ECLS-B.

32 In all tables described below, ECLS-B sample sizes are rounded to nearest 50 per the requirements of the National Center for Educational Statistics (NCES). Also, we use sampling weights that were provided with ECLS-B and the HSIS to account for the stratified sampling design, nonresponse, and a trimming factor. See Puma et al. (2010b) and the ECLS-B website.
<table>
<thead>
<tr>
<th></th>
<th>HSIS 3-Year-Olds</th>
<th>ECLS-B 3-Year-Olds</th>
<th>ECLS-B 4-Year-Olds</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Children female</td>
<td>.52 (.69)</td>
<td>.47 (.71)</td>
<td>.50 (.68)</td>
</tr>
<tr>
<td>% Children black</td>
<td>.29 (.61)</td>
<td>.35 (.63)</td>
<td>.18 (.54)</td>
</tr>
<tr>
<td>% Children Hispanic</td>
<td>.28 (.60)</td>
<td>.32 (.71)</td>
<td>.31 (.63)</td>
</tr>
<tr>
<td>% Children white</td>
<td>.23 (.55)</td>
<td>.26 (.69)</td>
<td>.45 (.71)</td>
</tr>
<tr>
<td>Age of child at baseline (months)</td>
<td>45.13* (5.38)</td>
<td>24.41 (1.71)</td>
<td>24.42 (1.65)</td>
</tr>
<tr>
<td>Age of child at assessment (months)</td>
<td>50.45* (5.29)</td>
<td>53.07 (6.88)</td>
<td>52.54 (6.57)</td>
</tr>
<tr>
<td>Years of maternal education</td>
<td>11.87 (2.55)</td>
<td>11.96 (2.59)</td>
<td>12.50 (3.09)</td>
</tr>
<tr>
<td>Age of mother</td>
<td>28.27 (7.86)</td>
<td>28.76 (8.81)</td>
<td>29.44 (8.09)</td>
</tr>
<tr>
<td>% Mothers married</td>
<td>.38 (.66)</td>
<td>.41 (.72)</td>
<td>.56 (.65)</td>
</tr>
<tr>
<td>% Mothers separated</td>
<td>.05 (.27)</td>
<td>.05 (.26)</td>
<td>.03 (.33)</td>
</tr>
<tr>
<td>% Mothers divorced</td>
<td>.05 (.27)</td>
<td>.06 (.35)</td>
<td>.06 (.36)</td>
</tr>
<tr>
<td>% Mothers never married</td>
<td>.34* (.63)</td>
<td>.45 (.72)</td>
<td>.32 (.65)</td>
</tr>
<tr>
<td>% Teenage mothers</td>
<td>.13 (.44)</td>
<td>.14 (.56)</td>
<td>.08 (.48)</td>
</tr>
<tr>
<td>Number of children under age 6</td>
<td>1.80 (1.12)</td>
<td>1.86 (1.12)</td>
<td>1.74 (1.06)</td>
</tr>
<tr>
<td>% Live in urban area</td>
<td>.79 (.56)</td>
<td>.83 (.51)</td>
<td>.82 (.49)</td>
</tr>
<tr>
<td>% Speak English at home</td>
<td>.77 (.57)</td>
<td>.77 (.60)</td>
<td>.77 (.58)</td>
</tr>
<tr>
<td>% Own house</td>
<td>.28 (.62)</td>
<td>.24 (.61)</td>
<td>.37 (.68)</td>
</tr>
<tr>
<td>N</td>
<td>2,449 500</td>
<td>3,500 1,993</td>
<td>3,300 200</td>
</tr>
</tbody>
</table>

* Statistically significant difference at the 5% level.

**NOTE.**—The F-test reports the p-value from a joint test of the equality of means between the HSIS and ECLS-B Head Start groups. Standard deviations are reported in parentheses. The test of significance is for the difference between the HSIS and the ECLS-B Head Start samples for the 3-year-olds and 4-year-olds.
This difference is consistent with the lower rates of marriage for the HSIS households. The proportion of mothers working is lower by 12–14 percentage points in the ECLS-B sample. The proportion of fathers (if present) working is the same among the 4-year-olds but lower among the HSIS 3-year-olds.

Table 3 replicates the experimental impact estimates derived from the HSIS subsamples. Both 3-year-olds and 4-year-olds in the treatment group...
have higher percentile scores in reading that are statistically significant at conventional levels, but there is no statistically significant difference in the mathematics percentile scores. There are no statistically significant differences in the health outcome measures. With regard to parenting behaviors, parents of 3-year-olds in Head Start are more likely to read at home

### Table 3

**HSIS Experimental Impacts**

<table>
<thead>
<tr>
<th>Intent to Treat(^a)</th>
<th>IV (LATE)(^b)</th>
<th>Treatment on the Treated(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-Year-Olds</td>
<td>4-Year-Olds</td>
<td>3-Year-Olds</td>
</tr>
<tr>
<td>Cognitive:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>4.28(^*)</td>
<td>4.56(^*)</td>
</tr>
<tr>
<td>(1.10)</td>
<td>(1.31)</td>
<td>(1.56)</td>
</tr>
<tr>
<td>Mathematics percentile</td>
<td>2.35</td>
<td>2.10</td>
</tr>
<tr>
<td>(1.80)</td>
<td>(1.51)</td>
<td>(2.60)</td>
</tr>
<tr>
<td>Health:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child health good/excellent</td>
<td>.05</td>
<td>-.03</td>
</tr>
<tr>
<td>(.03)</td>
<td>(.03)</td>
<td>(.04)</td>
</tr>
<tr>
<td>Health insurance</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>(.02)</td>
<td>(.02)</td>
<td>(.02)</td>
</tr>
<tr>
<td>Parenting:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>.07(^*)</td>
<td>.03</td>
</tr>
<tr>
<td>(.03)</td>
<td>(.03)</td>
<td>(.04)</td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>.03</td>
<td>.02</td>
</tr>
<tr>
<td>(.03)</td>
<td>(.03)</td>
<td>(.05)</td>
</tr>
<tr>
<td>Used time out</td>
<td>-.04</td>
<td>-.07(^*)</td>
</tr>
<tr>
<td>(.03)</td>
<td>(.03)</td>
<td>(.04)</td>
</tr>
<tr>
<td>Spanked child</td>
<td>-.07(^*)</td>
<td>.00</td>
</tr>
<tr>
<td>(.03)</td>
<td>(.03)</td>
<td>(.04)</td>
</tr>
<tr>
<td>Labor:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td>-104</td>
<td>196</td>
</tr>
<tr>
<td>(734)</td>
<td>(769)</td>
<td>(1,063)</td>
</tr>
<tr>
<td>Mother works</td>
<td>-.06</td>
<td>.01</td>
</tr>
<tr>
<td>(.03)</td>
<td>(.03)</td>
<td>(.04)</td>
</tr>
<tr>
<td>Father works (if present)</td>
<td>.05</td>
<td>.03</td>
</tr>
<tr>
<td>(.03)</td>
<td>(.03)</td>
<td>(.05)</td>
</tr>
</tbody>
</table>

**NOTE.**—Standard errors are reported in parentheses.

\(^{a}\) Reports mean difference based only on treatment assignment status.

\(^{b}\) Reports the average impact on those induced to change their Head Start status by the offer of treatment (the complier group).

\(^{c}\) Reports estimates from an IV procedure except reweighting the HSIS observations to have same propensity score distribution as ECLS-B participants.

\(^{*}\) Statistically significant difference from zero at the 5% significance level.
to their child and are less likely to discipline the child through spanking, consistent with Head Start’s parental education outreach. Parents of 4-year-olds are less likely to use time out. The Head Start program did not have any statistically significant impacts on the rates of parents working or on household income.

The HSIS has nontrivial fractions of “never-takers” (children assigned to treatment who never attended Head Start) and “always-takers” (children assigned to control who participated in Head Start); the percentage of never-takers was 16.3% among 3-year-olds and 22.3% among 4-year-olds, and the percentage of always-takers was 13.3% among 3-year-olds and 10.6% among 4-year-olds. For this reason, the first two columns in table 3 are interpretable as impacts of the offer of treatment or so-called intent-to-treat estimates. However, the treatment impacts we estimate in the ECLS-B use actual Head Start participation and estimate a treatment-on-the-treated parameter. To make the estimates more comparable, columns 3 and 4 of table 3 report IV estimates that instrument actual Head Start take-up in the HSIS with the randomized offer of treatment. These estimates are interpreted as impacts of Head Start for children induced to change their participation status by the treatment offer (the so-called LATE parameter, which is the average treatment effect for the “complier” group; see, e.g., Imbens and Angrist 1994).

Because the treated group in the ECLS-B sample is a mixture of compliers and always-takers, a potential source of discrepancy when comparing the ECLS-B and HSIS impacts is that treatments effects could differ for always-takers versus compliers. Table 4 reports covariate means for the following subgroups: always-takers, control group compliers, treatment group compliers, and never-takers. These group means can either be obtained directly in the data or inferred given the randomization of treatment (see, e.g., Abadie 2002; Kling, Liebman, and Katz 2007). Statistical tests reveal some statistically significant differences in covariate means between the always-taker and complier groups. As a way of controlling for potential differences in observables, we also report in table 3, in columns 5 and 6, an IV estimate that reweights the HSIS sample to have the same propensity score distribution as the ECLS-B. We interpret this estimate as an average impact of treatment on the treated (ATT) parameter analogous to what is estimated using the ECLS-B data.

A final important consideration, mentioned in Section II, is that the Head Start program draws a third of its participants from competing preschool programs that are in some cases close substitutes for Head Start and also receive public funding (Kline and Walters 2014). If the distribution of alternative care arrangements differs between the HSIS and ECLS-B samples, such differences might explain any observed differences in estimated impacts. However, we find that the distribution of alternative child care arrangements of HSIS control group (restricted to controls that do not receive Head Start) is very similar to the distribution in ECLS-B nonparticipant
sample, as shown in table 5. Among 3-year-olds in the HSIS control group, 31.1% were in center-based care, 44.2% were at home with one or more parent, and 24.7% were in “Other” care consisting of relative or nonrelative care in a home setting. For the 4-year-olds, the respective percentages were 41.0%, 41.6%, and 17.4%. Among ECLS-B Head Start nonparticipants at the third data collection round, 35.3% were in center-based care, 41.8% at home with parents, and 22.9% in relative or nonrelative care.  

A. Parametric Linear Model Estimates

Tables 6 and 7 show impact estimates derived from cross-section and difference-in-differences regression estimators for 3-year-olds and 4-year-olds, respectively. The first column of each table shows the ATT estimates from the experiment; the additional columns show the estimates derived from different nonexperimental estimators. Regressions with controls include all the variables listed in table 1, and regressions with lagged variables or difference-in-differences use the variables in table 2. To benchmark these parametric linear models appropriately with the matching models, we impose the same sample selection criterion and the same set of control variables in the regression models as used in the propensity score model (see below). We selected the set of control variables out of consideration of the Head Start eligibility criteria and to achieve a high classification rate in predicting program participation. Although this approach has no theoretical justification, Heckman, Ichimura, and Todd (1997), in their study of job training programs, showed that this approach produced a propensity score specification applicable to the HSIS and ECLS-B samples. In the HSIS, the exact variable construction is not given for confidentiality reasons but the focal care arrangement is described as “child care setting where the child spent a minimum of 5 hours between the hours of 8 a.m. and 6 p.m. Monday through Friday.” In the ECLS-B, they asked for the number of hours total spent per week in four different kinds of child care: center-based, Head Start, relative care, and nonrelative care. We assigned the child to Head Start based on the center director’s report. Conditional on not using Head Start, we assigned the child to a particular mode if the hours in that mode were the highest and if the child spent at least 5 hours in that mode (to make the question consistent with the HSIS). Otherwise the child was assigned to home. The distribution of choices in ECLS-B comes from the third round data collection round.

For the ECLS-B 3-year-olds, we only observe whether they were in Head Start but not the full distribution of child care choice. The choice distribution in table 5 refers to the choice distribution at the third round of the ECLS-B data collection. The ECLS-B distribution is skewed toward the distribution among HSIS 4-year-olds, which reflects that the ECLS-B children were slightly older at the third data round and that the use of center based care increases as children age. The questions are also asked differently in the HSIS and ECLS-B, which could contribute to some of the differences.
<table>
<thead>
<tr>
<th>Covariates</th>
<th>HSIS 3-Year-Olds</th>
<th></th>
<th>HSIS 4-Year-Olds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Treatment</td>
<td>Control</td>
<td>Treatment</td>
</tr>
<tr>
<td>% Children female</td>
<td>.50 (.04)</td>
<td>.52 (.02)</td>
<td>.47 (.03)</td>
<td>.53* (.04)</td>
</tr>
<tr>
<td>% Children black</td>
<td>.30 (.03)</td>
<td>.36* (.02)</td>
<td>.18* (.01)</td>
<td>.23 (.03)</td>
</tr>
<tr>
<td>% Children Hispanic</td>
<td>.27 (.02)</td>
<td>.31 (.02)</td>
<td>.22 (.01)</td>
<td>.49 (.05)</td>
</tr>
<tr>
<td>% Children white</td>
<td>.20 (.02)</td>
<td>.24 (.02)</td>
<td>.11* (.01)</td>
<td>.10 (.02)</td>
</tr>
<tr>
<td>Age of child at baseline (months)</td>
<td>45.20 (4.39)</td>
<td>45.21 (4.01)</td>
<td>45.69 (4.01)</td>
<td>56.12 (6.52)</td>
</tr>
<tr>
<td>Age of child at assessment (months)</td>
<td>50.35 (4.99)</td>
<td>50.68 (4.72)</td>
<td>50.62 (4.72)</td>
<td>61.58 (7.31)</td>
</tr>
<tr>
<td>Years of maternal education</td>
<td>11.70 (1.15)</td>
<td>11.87 (1.03)</td>
<td>11.81 (1.03)</td>
<td>11.44 (1.32)</td>
</tr>
<tr>
<td>Age of mother</td>
<td>28.34 (2.79)</td>
<td>28.33 (2.45)</td>
<td>27.98 (2.45)</td>
<td>28.81 (3.39)</td>
</tr>
<tr>
<td>% Mothers married</td>
<td>.36 (.03)</td>
<td>.39 (.02)</td>
<td>.25* (.02)</td>
<td>.39 (.04)</td>
</tr>
<tr>
<td></td>
<td>HSIS 3-Year-Olds</td>
<td>HSIS 4-Year-Olds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>Treatment</td>
<td>Control</td>
<td>Treatment</td>
</tr>
<tr>
<td>% Mothers separated</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.02*</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.00)</td>
</tr>
<tr>
<td>% Mothers divorced</td>
<td>.05</td>
<td>.05</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.00)</td>
</tr>
<tr>
<td>% Mothers never married</td>
<td>.34</td>
<td>.32</td>
<td>.42*</td>
<td>.22*</td>
</tr>
<tr>
<td></td>
<td>(.03)</td>
<td>(.02)</td>
<td>(.02)</td>
<td>(.01)</td>
</tr>
<tr>
<td>% Teenage mothers</td>
<td>.18</td>
<td>.19</td>
<td>.13*</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.02)</td>
<td>(.01)</td>
<td>(.01)</td>
</tr>
<tr>
<td>Number of children under age 6</td>
<td>1.70</td>
<td>1.83</td>
<td>1.76</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>(1.16)</td>
<td>(1.04)</td>
<td>(1.03)</td>
<td>(1.16)</td>
</tr>
<tr>
<td>% Live in urban area</td>
<td>.88</td>
<td>.80</td>
<td>.82</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>(.08)</td>
<td>(.02)</td>
<td>(.02)</td>
<td>(.06)</td>
</tr>
<tr>
<td>% Speak English at home</td>
<td>.72</td>
<td>.79</td>
<td>.77</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>(.06)</td>
<td>(.02)</td>
<td>(.02)</td>
<td>(.05)</td>
</tr>
<tr>
<td>% Own house</td>
<td>.21</td>
<td>.24</td>
<td>.32*</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.02)</td>
<td>(.02)</td>
<td>(.02)</td>
</tr>
</tbody>
</table>

Baseline outcomes:

|                                |        |        |        |        |        |        |        |        |
|Reading percentile              | 26.17  | 28.88  | 31.36  | 34.35* |        |        |        |        |
|                                | (2.86) | (.96)  | (.73)  | (2.12) |        |        |        |        |
|Mathematics percentile          | 30.13  | 33.77  | 35.89  | 34.32  |        |        |        |        |
|                                | (4.22) | (1.48) | (1.22) | (2.43) |        |        |        |        |
|Child health                    | .82    | .79    | .78    | .83    | .73    | .78    | .81    | .78     |         |           |         |           |
|                                | (.07)  | (.02)  | (.01)  | (.05)  | (.08)  | (.02)  | (.02)  | (.05)   |         |           |         |           |
|Health insurance                | .88    | .87    | .91    | .90    | .93    | .86    | .81    | .91     |         |           |         |           |
|                                | (.08)  | (.01)  | (.01)  | (.06)  | (.10)  | (.02)  | (.01)  | (.06)   |         |           |         |           |
|Read to child                   | .27    | .35*   | .37*   | .29    | .38    | .35    | .36    | .31     |         |           |         |           |
|                                | (.02)  | (.02)  | (.02)  | (.02)  | (.04)  | (.03)  | (.02)  | (.02)   |         |           |         |           |
Table 4 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>HSIS 3-Year-Olds</th>
<th></th>
<th>HSIS 4-Year-Olds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Treatment</td>
<td>Control</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td>Always-Takers</td>
<td>Compliers</td>
<td>Never-Takers</td>
<td>Always-Takers</td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>3.75 (0.36)</td>
<td>3.73 (0.03)</td>
<td>3.73 (0.02)</td>
<td>3.87 (0.34)</td>
</tr>
<tr>
<td>Used time out</td>
<td>0.66 (.06)</td>
<td>0.65 (.06)</td>
<td>0.60 (.02)</td>
<td>0.66 (.04)</td>
</tr>
<tr>
<td>Spanked child</td>
<td>0.42 (.04)</td>
<td>0.51* (.02)</td>
<td>0.46 (.02)</td>
<td>0.48 (.04)</td>
</tr>
<tr>
<td>Household income</td>
<td>17,923 (1,820)</td>
<td>15,952 (700)</td>
<td>16,134 (574)</td>
<td>18,975 (1,754)</td>
</tr>
<tr>
<td>Mother works</td>
<td>0.51 (.05)</td>
<td>0.49 (.03)</td>
<td>0.50 (.02)</td>
<td>0.53 (.05)</td>
</tr>
<tr>
<td>Father works</td>
<td>0.85 (.12)</td>
<td>0.78 (.03)</td>
<td>0.83 (.02)</td>
<td>0.83 (.10)</td>
</tr>
</tbody>
</table>

NOTE.—Standard errors are reported in parentheses. Statistical tests compare always-takers to the two complier groups (control and treatment) and to the never-takers, separately for 3-year-olds and 4-year-olds.

* Indicates statistically significant difference at the 5% level from the always-takers.
that came close to replicating experimental results. It is possible that using a different set of conditioning variables or even a smaller set could lead to better performance for certain outcomes. However, given the large number of estimators and outcomes analyzed in this paper, we maintain a consistent set of conditioning variables throughout.

The difference-in-differences estimates are not available for the 3-year-old subsample because there are no baseline scores (the children were too young to be administered the test). We report two measures of the difference between the impact estimates in the HSIS and the ECLS-B. The first is the percent difference from the HSIS impact estimate (%Δ from HSIS Impact), and the second is the difference from the HSIS impact in standard deviation units (Δ from HSIS Impact (σ)). Both measures allow comparisons to be made across outcome domains.

As seen in the third column of table 6, the cross-section estimates (the only available estimates for 3-year-olds for test scores) come close to replicating the experimental results. With regard to health, under the experiment, there were no statistically significant differences in child health or in the proportion of children covered by health insurance, which is generally consistent with all of the nonexperimental estimates. However, the nonexperimental estimators do not reproduce the positive impacts on the percentage of parents who read to the child or on parenting practices (time out, spanking) that are observed under the experiment. All of the regression models indicate negative effects on the program on household income, which was not observed under the experiment, and one model (difference-in-differences with controls) shows a negative impact on father working that was also not observed under the experiment. Joint F-tests of the equal-

Table 5
Distribution of Child Care Choices

<table>
<thead>
<tr>
<th></th>
<th>Head Start</th>
<th>Center</th>
<th>Home</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSIS treatment (participants): 3-year-olds</td>
<td>100.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td></td>
<td>4-year-olds</td>
<td>100.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>HSIS control (nonparticipants): 3-year-olds</td>
<td>.0</td>
<td>31.1</td>
<td>44.2</td>
<td>24.7</td>
</tr>
<tr>
<td></td>
<td>4-year-olds</td>
<td>.0</td>
<td>41.0</td>
<td>41.6</td>
</tr>
<tr>
<td>ECLS-B treatment (participants): 3-year-olds</td>
<td>100.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td></td>
<td>4-year-olds</td>
<td>.0</td>
<td>35.3</td>
<td>41.8</td>
</tr>
</tbody>
</table>

NOTE.—“Home” refers to care given in the home by a parent. “Other” includes both relative and nonrelative child care in and outside of the child’s home.

35 Child health is reported as excellent, very good, good, fair, or poor. Our measure assigns a 1 if reported to be excellent or very good, otherwise a 0.
Table 6
Regression-Based Estimators for 3-Year-Olds

<table>
<thead>
<tr>
<th></th>
<th>ECLS-B Nonexperimental Head Start Impact Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HSIS Impact</td>
</tr>
<tr>
<td>Cognitive:</td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>5.31*</td>
</tr>
<tr>
<td></td>
<td>(1.86)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-87</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>-2.47</td>
</tr>
<tr>
<td>Mathematics percentile</td>
<td>.86</td>
</tr>
<tr>
<td>(3.18)</td>
<td>(2.07)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-41</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>-1.11</td>
</tr>
<tr>
<td>F-test</td>
<td>.63</td>
</tr>
<tr>
<td>Health:</td>
<td></td>
</tr>
<tr>
<td>Child health good/excellent</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-469</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>-4.9</td>
</tr>
<tr>
<td>Health insurance</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>51</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>.52</td>
</tr>
<tr>
<td>F-test</td>
<td>1.00</td>
</tr>
<tr>
<td>Parenting:</td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>.11*</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
</tr>
<tr>
<td>%Δ from HSIS Impact</td>
<td>-.150</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>-3.40</td>
</tr>
<tr>
<td>Domain</td>
<td>Mean Difference (Regression with Controls)</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>.09</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>−319</td>
</tr>
<tr>
<td>∆ from HSIS impact (a)</td>
<td>−4.17</td>
</tr>
<tr>
<td>Used time out</td>
<td>−.01</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>−130</td>
</tr>
<tr>
<td>Spanked child</td>
<td>−.10</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>130</td>
</tr>
<tr>
<td>∆ from HSIS impact (a)</td>
<td>2.43</td>
</tr>
<tr>
<td>Labor:</td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td>−485</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>−1,877</td>
</tr>
<tr>
<td>∆ from HSIS impact (a)</td>
<td>−6.65</td>
</tr>
<tr>
<td>Mother works</td>
<td>−.03</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>7</td>
</tr>
<tr>
<td>∆ from HSIS impact (a)</td>
<td>.03</td>
</tr>
<tr>
<td>Father works (if present)</td>
<td>.04</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>−251</td>
</tr>
<tr>
<td>∆ from HSIS impact (a)</td>
<td>−1.81</td>
</tr>
<tr>
<td>F-test</td>
<td>.05</td>
</tr>
<tr>
<td>F-test overall</td>
<td>.02</td>
</tr>
</tbody>
</table>

**NOTE.** — *F*-tests are reported for each domain and overall, which gives the *p*-value for an omnibus test of the equality of the HSIS impact and the nonexperimental impacts estimated in the ECLS-B. The variance-covariance matrix of the impact estimates across outcomes by model was estimated by bootstrapping. Standard errors are reported in parentheses.

* Control variables include all the variables listed in table 1.

* Statistically significant from zero at the 5% significance level.

† Statistically significant difference from the HSIS impact at the 5% significance level.

ity of the experimental and nonexperimental impacts by domain and model generally do not reject the null of equality. However, overall $F$-tests reject the null for three of the estimators.

Table 7 shows analogous results for 4-year-olds. In the cognitive domain, the difference-in-differences estimator with controls results come close to replicating the experimental impacts. The nonexperimental estimators generally reproduce the experimental result of no effect on child health or on parenting. Again, some nonexperimental estimators suggest a negative effect of Head Start participation on income, which was not observed under the experiment. For the 4-year-olds, the $F$-tests by domain and overall both do not reject the null, which could be a consequence of the smaller sample size for the ECLS-B 4-year-olds.

The nonexperimental regression estimators sometimes produce estimates that are very close to the experimental estimates, but their performance is not consistent. The estimators are fairly reliable for estimating cognitive test score effects, with the difference-in-differences and the value-added regressions being among the best in reproducing the experimental test score effects. They also generally reproduce the finding of small and statistically insignificant impacts on child health, but they miss the finding of a significant effect on parents reading to the child for 3-year-olds. The estimators tend to exhibit large biases for the household income outcome measure.

B. Matching Estimates

We next implement propensity score matching estimators, where the propensity scores are estimated by a probit model using data on ECLS-B participants and nonparticipants. In specifying the propensity score model, we need to consider the process by which parents enroll their children in Head Start. Parents would be expected to weigh the expected benefits of participating against the expected costs. Benefits might include cognitive test score gains, potential increases in socioemotional development, nutritional inputs provided at the center, the child’s enjoyment, and the time freed up for the mother to engage in other activities, such as work. Expected costs would include transportation costs to travel to and from the center, as well as any direct costs of participation (in terms of time or money). If there are available alternatives to Head Start, parents would choose Head Start when the benefit-cost calculation finds Head Start to be maximal among all alternatives; therefore, characteristics of alternatives might figure into the decision. Also, parents who expect they would not be eligible for Head Start would not participate (they might not apply or they might be rejected upon application).

We include in the propensity score model three types of variables: variables that forecast relative benefits of Head Start participation, variables that are likely cost determinants, and variables that capture program eligibility. Specifically, we include child characteristics (gender, age, race/ethnicity),
mother’s characteristics (education level, marital status, teen motherhood, presence of other young children in the household, English-speaking ability), geographic characteristics (% urban), and family income. An important determinant of cost is distance to the nearest Head Start center and distance to alternative centers, but such variables are not available in these data. These variables constitute an important source of unobserved variation in participation after conditioning on observables.

Initially, we estimated a propensity score model with all variables from table 1. Balancing tests were performed by regressing the covariates on a power series in the propensity scores and the propensity scores interacted with a participation indicator (see Smith and Todd 2005b). The tests revealed that the covariates were mostly balanced except for imbalances by income, teenage motherhood, and owning a house. This is perhaps not surprising given that Head Start programs can make individual eligibility judgments taking into account other indicators of finances (e.g., owning a home) or personal situations (e.g., being a teenage mother).

To address this imbalance, we made three modifications. First, we restricted the ECLS-B Non–Head Start comparison sample to those reporting household income less than $50,000. Although Head Start does allow some children with higher family income (up to 35% can be from families with income less than 135% of the poverty line and up to 10% of children can be from families with incomes above the poverty thresholds), the $50,000 cutoff is at the 98th percentile of household income in the HSIS. Those with family income above $50,000 are highly unlikely to be eligible nonparticipants. Second, we estimate the model including indicator variables for household income falling into bins of $2,500. This created balance in the income variable, but it still left imbalances by teenage motherhood and owning a house. To remedy the remaining imbalance, we interacted the teenage mother indicator with income (which ensured balance for this variable) and interacted an indicator for owning a home with income and income squared. These additions serve to balance the sample for every variable dimension using the regression balancing tests implemented in Smith and Todd (2005b).

Table 8 shows the probit propensity score estimates. The child being white, the mother being married, the mother being never married, living in an urban area, or owning a home all decrease the probability of participation. The estimated coefficients on the income bin indicators (not reported) show that lower income is associated with a higher likelihood of participation, consistent with Head Start program eligibility rules. The percent correctly classified using the hit or miss method with the sample proportion in Head Start as the cutoff is 64.5%. Table 9 shows the p-values from the regression balancing tests.

Figure 1 plots the propensity score distributions, providing a visual depiction of the degree of overlap between ECLS-B Head Start participants and nonparticipants used for the matching analysis. In the first panel, the histogram for the ECLS-B Head Start participants (with a white fill) is over-
Table 7
Regression-Based Estimators for 4-Year-Olds

<table>
<thead>
<tr>
<th></th>
<th>HSIS Impact</th>
<th>Mean Difference</th>
<th>Regression with Controls</th>
<th>Difference-in-Differences with Controls</th>
<th>Differences-in-Differences with Controls</th>
<th>Lagged Value-Added</th>
<th>Lagged Value-Added with Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cognitive:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>7.23*</td>
<td>-2.08†</td>
<td>4.28</td>
<td>5.75</td>
<td>6.26</td>
<td>2.57</td>
<td>5.80</td>
</tr>
<tr>
<td></td>
<td>(2.18)</td>
<td>(2.93)</td>
<td>(3.18)</td>
<td>(2.78)</td>
<td>(2.77)</td>
<td>(2.58)</td>
<td>(2.81)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-4.26</td>
<td>-1.35</td>
<td>-0.68</td>
<td>-0.44</td>
<td>-2.13</td>
<td>-6.6</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (σ)</td>
<td>-184</td>
<td>-58</td>
<td>-5</td>
<td>5</td>
<td>-60</td>
<td>-10</td>
<td></td>
</tr>
<tr>
<td>Mathematics percentile</td>
<td>4.32</td>
<td>-3.62†</td>
<td>1.83</td>
<td>4.12</td>
<td>4.53</td>
<td>1.72</td>
<td>3.91</td>
</tr>
<tr>
<td></td>
<td>(2.54)</td>
<td>(2.52)</td>
<td>(2.86)</td>
<td>(2.33)</td>
<td>(2.59)</td>
<td>(1.95)</td>
<td>(2.31)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-3.12</td>
<td>-0.98</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-1.02</td>
<td>-1.16</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (σ)</td>
<td>-3.12</td>
<td>-0.98</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-1.02</td>
<td>-1.16</td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td>.41</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.90</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td><strong>Health:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child health good/excellent</td>
<td>.00</td>
<td>-.00</td>
<td>.01</td>
<td>.05</td>
<td>.04</td>
<td>.01</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.04)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.04)</td>
<td>(.04)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-172</td>
<td>270</td>
<td>2,209</td>
<td>2,041</td>
<td>549</td>
<td>957</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (σ)</td>
<td>-0.07</td>
<td>.12</td>
<td>.96</td>
<td>.89</td>
<td>.24</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>.01</td>
<td>.03</td>
<td>.07</td>
<td>-.01</td>
<td>.04</td>
<td>.02</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>(.03)</td>
<td>(.02)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.02)</td>
<td>(.02)</td>
<td>(.01)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>130</td>
<td>450</td>
<td>-176</td>
<td>210</td>
<td>33</td>
<td>334</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (σ)</td>
<td>.50</td>
<td>1.72</td>
<td>-.67</td>
<td>.80</td>
<td>.12</td>
<td>1.27</td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td>1.00</td>
<td>.81</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.94</td>
<td></td>
</tr>
<tr>
<td><strong>Parenting:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>.07</td>
<td>.01</td>
<td>.05</td>
<td>-.01</td>
<td>-.06</td>
<td>.00</td>
<td>-.01</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-81</td>
<td>-26</td>
<td>-115</td>
<td>-185</td>
<td>-95</td>
<td>-116</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (σ)</td>
<td>-1.03</td>
<td>-.33</td>
<td>-1.46</td>
<td>-2.35</td>
<td>-1.21</td>
<td>-1.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>.02</td>
<td>.03</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>.01</td>
<td>.02</td>
<td>.03</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-184</td>
<td>-252</td>
<td>-812</td>
<td>-1051</td>
<td>-494</td>
<td>-868</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>-53</td>
<td>-72</td>
<td>-2.32</td>
<td>-3.00</td>
<td>-1.41</td>
<td>-2.48</td>
<td></td>
</tr>
<tr>
<td>Used time out</td>
<td>-.10</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>164</td>
<td>164</td>
<td>165</td>
<td>99</td>
<td>164</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>3.05</td>
<td>3.04</td>
<td>3.06</td>
<td>1.85</td>
<td>3.05</td>
<td>2.38</td>
<td></td>
</tr>
<tr>
<td>Spanked child</td>
<td>-.08</td>
<td>.04</td>
<td>.03</td>
<td>.05</td>
<td>.08</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>157</td>
<td>135</td>
<td>172</td>
<td>201</td>
<td>162</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>2.01</td>
<td>1.73</td>
<td>2.20</td>
<td>2.57</td>
<td>2.08</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td>.70</td>
<td>.85</td>
<td>.59</td>
<td>.58</td>
<td>.58</td>
<td>.70</td>
<td></td>
</tr>
</tbody>
</table>

**Labor:**

<table>
<thead>
<tr>
<th></th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household income</td>
<td>501</td>
<td>1,589</td>
<td>-312</td>
<td>2,541</td>
<td>1,144</td>
<td>571</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-1,401</td>
<td>-497</td>
<td>-1,277</td>
<td>-599</td>
<td>-1,263</td>
<td>-608</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>-5.28</td>
<td>-1.87</td>
<td>-4.82</td>
<td>-2.26</td>
<td>-4.77</td>
<td>-2.29</td>
</tr>
<tr>
<td>Mother works</td>
<td>.00</td>
<td>.06</td>
<td>.01</td>
<td>.11</td>
<td>.05</td>
<td>.03</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-1,589</td>
<td>-312</td>
<td>2,541</td>
<td>1,144</td>
<td>571</td>
<td>-109</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>-1.13</td>
<td>-22</td>
<td>1.80</td>
<td>.81</td>
<td>.40</td>
<td>.08</td>
</tr>
<tr>
<td>Father works (if present)</td>
<td>.08</td>
<td>.03</td>
<td>.02</td>
<td>.08</td>
<td>.00</td>
<td>.02</td>
</tr>
<tr>
<td>%Δ from HSIS impact</td>
<td>-1.98</td>
<td>-1.91</td>
<td>.11</td>
<td>-1.53</td>
<td>-1.07</td>
<td>-2.08</td>
</tr>
<tr>
<td>Δ from HSIS impact (α)</td>
<td>.30</td>
<td>.95</td>
<td>.46</td>
<td>.87</td>
<td>.43</td>
<td>.80</td>
</tr>
<tr>
<td>F-test</td>
<td>.51</td>
<td>.99</td>
<td>.90</td>
<td>.97</td>
<td>.85</td>
<td>.97</td>
</tr>
</tbody>
</table>

**NOTE.**—F-tests are reported for each domain and overall, which gives the p-value for an omnibus test of the equality of the HSIS impact and the nonexperimental impacts estimated in the ECLS-B. The variance-covariance matrix of the impact estimates across outcomes by model was estimated by bootstrapping. Standard errors are reported in parentheses.

* Control variables include all the variables listed in table 1.

* Statistically significant difference from zero at the 5% significance level.

* Statistically significant difference from the HSIS impact at the 5% significance level.
Table 8
Probit Propensity Score Model

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Marginal Effects (Average Derivatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child is female</td>
<td>-.0856</td>
<td>-.0185</td>
</tr>
<tr>
<td></td>
<td>(.080)</td>
<td>(.017)</td>
</tr>
<tr>
<td>Child is black</td>
<td>.1900</td>
<td>.0410</td>
</tr>
<tr>
<td></td>
<td>(.160)</td>
<td>(.034)</td>
</tr>
<tr>
<td>Child is Hispanic</td>
<td>-.0078</td>
<td>-.0017</td>
</tr>
<tr>
<td></td>
<td>(.161)</td>
<td>(.035)</td>
</tr>
<tr>
<td>Child is white</td>
<td>-.3277*</td>
<td>-.0707*</td>
</tr>
<tr>
<td></td>
<td>(.164)</td>
<td>(.035)</td>
</tr>
<tr>
<td>Child’s age at baseline</td>
<td>-.0065</td>
<td>-.0014</td>
</tr>
<tr>
<td></td>
<td>(.020)</td>
<td>(.004)</td>
</tr>
<tr>
<td>Child’s age at assessment</td>
<td>.0065</td>
<td>.0014</td>
</tr>
<tr>
<td></td>
<td>(.021)</td>
<td>(.005)</td>
</tr>
<tr>
<td>Mother’s years of education</td>
<td>.0023</td>
<td>.0005</td>
</tr>
<tr>
<td></td>
<td>(.018)</td>
<td>(.004)</td>
</tr>
<tr>
<td>Age of mother</td>
<td>.0054</td>
<td>.0012</td>
</tr>
<tr>
<td></td>
<td>(.008)</td>
<td>(.002)</td>
</tr>
<tr>
<td>Mother is married</td>
<td>-.4222*</td>
<td>-.0911*</td>
</tr>
<tr>
<td></td>
<td>(.250)</td>
<td>(.054)</td>
</tr>
<tr>
<td>Mother is separated</td>
<td>-.3842</td>
<td>-.0829</td>
</tr>
<tr>
<td></td>
<td>(.318)</td>
<td>(.069)</td>
</tr>
<tr>
<td>Mother is divorced</td>
<td>-.4472</td>
<td>-.0965</td>
</tr>
<tr>
<td></td>
<td>(.290)</td>
<td>(.063)</td>
</tr>
<tr>
<td>Mother is never married</td>
<td>-.4432*</td>
<td>-.0956*</td>
</tr>
<tr>
<td></td>
<td>(.251)</td>
<td>(.054)</td>
</tr>
<tr>
<td>Teenage mother</td>
<td>-.0516</td>
<td>-.0111</td>
</tr>
<tr>
<td></td>
<td>(.253)</td>
<td>(.055)</td>
</tr>
<tr>
<td>Number of children under age 6</td>
<td>.0487</td>
<td>.0105</td>
</tr>
<tr>
<td></td>
<td>(.048)</td>
<td>(.010)</td>
</tr>
<tr>
<td>Urban area</td>
<td>-.2798*</td>
<td>-.0604*</td>
</tr>
<tr>
<td></td>
<td>(.117)</td>
<td>(.025)</td>
</tr>
<tr>
<td>English is primary language</td>
<td>-.0441</td>
<td>-.0095</td>
</tr>
<tr>
<td></td>
<td>(.125)</td>
<td>(.027)</td>
</tr>
<tr>
<td>Own house</td>
<td>-.7092*</td>
<td>-.1530*</td>
</tr>
<tr>
<td></td>
<td>(.419)</td>
<td>(.090)</td>
</tr>
<tr>
<td>Constant</td>
<td>-.3770</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.866)</td>
<td></td>
</tr>
</tbody>
</table>

**NOTE.**—Dependent variable is Head Start participation. Additional controls include indicators for baseline income falling in bins of $2,500, an indicator for “Teenage mother” interacted with “Baseline income,” and an indicator for “Own house” interacted with “Baseline income” and “Baseline income squared.” Standard errors are reported in parentheses.

* Statistically significant difference from zero at the 5% significance level.
laid on the histogram ECLS-B Head Start nonparticipants (with a light-gray fill). One hundred percent of the data are in the overlap region, so there is no failure of common support for these data. We also used the propensity score model to impute propensity scores for the HSIS children. In the second panel, we overlay the histogram for ECLS-B Head Start participants on the histogram for the HSIS data (with a dark-gray fill). The range of the propensity scores is similar in the ECLS-B data and in the HSIS data, and the shapes of distributions are also similar.

1. Cross-Section Matching

As previously discussed, the cross-section matching estimator imposes a conditional independence assumption (CIA). This implies that the decision to participate in Head Start does not depend directly on child outcomes. Such an assumption is plausible if parents do not have perfect knowledge about the gains from their child participating in the program and instead form a forecast of gains based on observable characteristics of the child (e.g., age, race/ethnicity/gender).

Table 10 shows the impact estimates derived from cross-section local linear matching, where the bandwidth is chosen using a cross-validation

<table>
<thead>
<tr>
<th>Table 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propensity Score Balancing Tests</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>$p$-Value</td>
</tr>
<tr>
<td>Child is female</td>
</tr>
<tr>
<td>Child is black</td>
</tr>
<tr>
<td>Child is Hispanic</td>
</tr>
<tr>
<td>Child is white</td>
</tr>
<tr>
<td>Child’s age at baseline</td>
</tr>
<tr>
<td>Child’s age at assessment</td>
</tr>
<tr>
<td>Mother’s years of education</td>
</tr>
<tr>
<td>Age of mother</td>
</tr>
<tr>
<td>Mother is married</td>
</tr>
<tr>
<td>Mother is separated</td>
</tr>
<tr>
<td>Mother is divorced</td>
</tr>
<tr>
<td>Mother never married</td>
</tr>
<tr>
<td>Teenage mother</td>
</tr>
<tr>
<td>Number children under age 6</td>
</tr>
<tr>
<td>Urban area</td>
</tr>
<tr>
<td>English is primary language</td>
</tr>
<tr>
<td>Own house</td>
</tr>
<tr>
<td>Household income baseline</td>
</tr>
</tbody>
</table>

**Note.** For each variable $Z_k$, we estimated $Z_k = \beta_0 + \beta_1 P(Z) + \beta_2 P(Z)^2 + \beta_3 P(Z)^3 + \beta_4 P(Z)^4 + \beta_5 DP(Z) + \beta_6 DP(Z)^2 + \beta_7 DP(Z)^3 + \beta_8 DP(Z)^4 + \eta$, where $D$ is an indicator that equals 1 if the child attended Head Start and 0 otherwise. The $p$-values are from an $F$-test on the joint null that the coefficients on the terms interacted with a Head Start indicator equal zero.
Fig. 1.—Propensity score histograms for common covariates in HSIS and ECLS-B.
## Table 10
Cross-Section Local Linear Matching Estimators

<table>
<thead>
<tr>
<th></th>
<th>3-Year-Olds</th>
<th></th>
<th>4-Year-Olds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head Start Impact</td>
<td>Δ from HSIS</td>
<td></td>
<td>Head Start Impact</td>
</tr>
<tr>
<td></td>
<td>HSIS ECLS-B %</td>
<td>σ</td>
<td>HSIS ECLS-B %</td>
<td>σ</td>
</tr>
<tr>
<td>Cognitive:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>5.31$^*$</td>
<td>2.92</td>
<td>−45</td>
<td>−1.29</td>
</tr>
<tr>
<td></td>
<td>(1.86)</td>
<td>(2.37)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics percentile</td>
<td>0.86</td>
<td>2.07</td>
<td>141</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>(3.18)</td>
<td>(2.19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td>.93</td>
<td></td>
<td>.76</td>
<td></td>
</tr>
<tr>
<td>Health:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child health good/excellent</td>
<td>0.00</td>
<td>0.03</td>
<td>639</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td>(.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>0.02</td>
<td>0.03</td>
<td>5</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td>.99</td>
<td></td>
<td>.99</td>
<td></td>
</tr>
<tr>
<td>Parenting:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>.11$^*$</td>
<td>−.05$^*$</td>
<td>−141</td>
<td>−3.20</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>.09</td>
<td>−.06</td>
<td>−175</td>
<td>−2.29</td>
</tr>
<tr>
<td></td>
<td>(.07)</td>
<td>(.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used time out</td>
<td>−.01</td>
<td>.02</td>
<td>434</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanked child</td>
<td>−.10</td>
<td>.01</td>
<td>110</td>
<td>2.05</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td>.36</td>
<td></td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>Labor:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td>−485</td>
<td>−8,558$^*$</td>
<td>−1,665</td>
<td>−5.90</td>
</tr>
<tr>
<td></td>
<td>(1,369)</td>
<td>(1,577)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother works</td>
<td>−.03</td>
<td>−.01</td>
<td>71</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father works (if present)</td>
<td>.04</td>
<td>−.08</td>
<td>−275</td>
<td>−1.99</td>
</tr>
<tr>
<td></td>
<td>(.06)</td>
<td>(.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td>.12</td>
<td></td>
<td>.74</td>
<td></td>
</tr>
<tr>
<td>F-test overall</td>
<td>.19</td>
<td></td>
<td>.96</td>
<td></td>
</tr>
</tbody>
</table>

**Note.**—F-tests are reported for each domain and overall, which gives the p-value for an omnibus test of the equality of the HSIS impact and the nonexperimental impacts estimated in the ECLS-B. The variance-covariance matrix of the impact estimates across outcomes by model was estimated by bootstrapping. Standard errors are reported in parentheses.

* For the HSIS impact, denotes a statistically significant difference from zero at the 5% significance level.

$^*$ For the nonexperimental impact, denotes a statistically significant difference between the HSIS impact at the 5% significance level.
criterion. The biases associated with the cross-section matching estimator are quite similar to those observed for the parametric regression estimators. The matching estimates of program impacts on reading scores are similar to the experimental estimates for the 3-year-old subsample but not for the 4-year-old subsample. Matching reproduces the finding based on the experimental data of no impacts on health outcomes. However, the matching estimates also suggest no impacts on parental behaviors, whereas the HSIS data indicate an impact for 3-year-olds on parents reading to the child. Again, the nonexperimental estimators find a negative effect on income that was not observed in the experiment. The level of bias is sometimes substantial and would imply different conclusions about program impacts.

2. Difference-in-Differences Matching

The DID matching estimator imposes a bias stability assumption (BSA). As previously noted, BSA allows the Head Start participation decision to depend on time-invariant unobservables. If families do not change geographic location over the time period covered by our data, then distance to a Head Start center or to other centers might be considered time-invariant unobservables that are relevant to participation decisions, as might be the proximity of extended family members who are able to provide alternative child care arrangements.

Table 11 shows the matching results for difference-in-differences matching. It is not possible to have difference-in-differences matching estimators for 3-year-old cognitive test scores because no test was given at baseline. For the 4-year-olds in the cognitive domain, the nonexperimental impacts are remarkably close and are within 2%–13% (0.05σ–0.23σ) of the experimental impacts. The difference-in-differences estimator also generally replicates the experimental findings of no impacts on health or parenting behaviors, although it again misses the increase in parents reading to the child for 3-year-olds. With one exception, the difference-in-differences estimates indicate no statistically significant impact on household labor variables, as was also found under the experiment.

C. Specification Tests

As discussed in Section I, Heckman and Hotz (1989) found that when they applied two types of specification tests to the NSW data, they were able to rule out estimators with the largest biases. In this section, we apply one of the specification tests that they used to the regression, cross-section, and difference-in-differences matching estimators from tables 6, 7, 10, and 11.

The cross-validated bandwidths for the cross-section and difference-in-differences estimators are displayed in table 14. We also estimated the models (in results not presented here) using the rule-of-thumb of Silverman (1986). The pattern of results was not sensitive to the method of choosing the bandwidth.
## Table 11
### Difference-in-Differences Local Linear Matching Estimators

<table>
<thead>
<tr>
<th></th>
<th>3-Year-Olds</th>
<th></th>
<th>4-Year-Olds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head Start</td>
<td>Δ from HSIS</td>
<td>Head Start</td>
<td>Δ from HSIS</td>
</tr>
<tr>
<td></td>
<td>Impact</td>
<td>HSIS ECLS-B</td>
<td>Impact</td>
<td>HSIS ECLS-B</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>σ</td>
<td>%</td>
<td>σ</td>
</tr>
<tr>
<td><strong>Cognitive:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>7.23</td>
<td>7.35</td>
<td>2</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>(2.18)</td>
<td>(3.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>4.32</td>
<td>4.90</td>
<td>13</td>
<td>.23</td>
</tr>
<tr>
<td>percentile</td>
<td>(2.54)</td>
<td>(2.53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F-test</strong></td>
<td>.99</td>
<td></td>
<td>.99</td>
<td></td>
</tr>
<tr>
<td><strong>Health:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child health good/</td>
<td>.00</td>
<td>.02</td>
<td>754</td>
<td>.33</td>
</tr>
<tr>
<td>excellent</td>
<td>(.04)</td>
<td>(.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>.02</td>
<td>.01</td>
<td>.01</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F-test</strong></td>
<td>.97</td>
<td>.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Parenting:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>.11*</td>
<td>-.03†</td>
<td>-127</td>
<td>-2.88</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.03)</td>
<td>(.05)</td>
<td>(.05)</td>
</tr>
<tr>
<td>Parental safety</td>
<td>.09</td>
<td>-.07</td>
<td>-184</td>
<td>-2.40</td>
</tr>
<tr>
<td>practices</td>
<td>(.07)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.08)</td>
</tr>
<tr>
<td>Used time out</td>
<td>-.01</td>
<td>.00</td>
<td>138</td>
<td>.19</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.06)</td>
<td>(.08)</td>
</tr>
<tr>
<td>Spanked child</td>
<td>-.10</td>
<td>.01</td>
<td>115</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.06)</td>
<td>(.06)</td>
</tr>
<tr>
<td><strong>F-test</strong></td>
<td>.36</td>
<td>.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Labor:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td>-485</td>
<td>-4830†</td>
<td>501</td>
<td>-2,959</td>
</tr>
<tr>
<td></td>
<td>(1,369)</td>
<td>(1,081)</td>
<td>(1,329)</td>
<td>(1,407)</td>
</tr>
<tr>
<td>Mother works</td>
<td>-.03</td>
<td>.06</td>
<td>317</td>
<td>1.54</td>
</tr>
<tr>
<td>(if present)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.06)</td>
<td>(.08)</td>
</tr>
<tr>
<td>Father works</td>
<td>.04</td>
<td>-.12</td>
<td>-383</td>
<td>-2.77</td>
</tr>
<tr>
<td>(if present)</td>
<td>(.06)</td>
<td>(.07)</td>
<td>(.05)</td>
<td>(.05)</td>
</tr>
<tr>
<td><strong>F-test</strong></td>
<td>.24</td>
<td>.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F-test overall</strong></td>
<td>.22</td>
<td>.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.**—F-tests are reported for each domain and overall, which gives the p-value for an omnibus test of the equality of the HSIS impact and the nonexperimental impacts estimated in the ECLS-B. The variance-covariance matrix of the impact estimates across outcomes by model was estimated by bootstrapping. Standard errors are reported in parentheses.

* For the HSIS impact, denotes a statistically significant difference from zero at the 5% significance level.

† For the nonexperimental impact estimates, denotes a statistically significant difference between the HSIS impact at the 5% significance level.
The test is the pre-program alignment test, in which each candidate estimator is applied to outcome data from a period prior to the program (i.e., to random assignment). For the ECLS-B 3-year-olds, we apply the cross-section Heckman-Hotz tests using outcome data from the second round. For the ECLS-B 4-year-olds, we apply the cross-section Heckman-Hotz tests using data from the third round and difference-in-differences Heckman-Hotz tests using data from the second and third rounds. Because of missing survey instruments in earlier rounds, not all of the Heckman-Hotz tests are available, especially for the 3-year-old cohort. For example, math and reading scores are not available prior to the third round, and parental behaviors are not available in the first round. This test actually tests the joint null that the outcome and participation processes are similar in the pre-program and post-program periods and that the estimator eliminates bias. 37

In table 12, we report Heckman-Hotz tests for the regression-based estimators of tables 6 and 7. Because the 3-year-olds do not have pre-baseline data for any of the outcomes, we simply report the regression with controls specification on the baseline outcomes for that subgroup. Few of the estimates are statistically significantly different from zero. Table 13 reports analogous pre-program test results for both the cross-section and difference-in-difference local linear matching estimators. Again for the tests that use lagged outcomes, the test can only be implemented using the older 4-year-olds. In the cognitive domain, the Heckman-Hotz test rejects the cross-section matching estimator for 4-year-olds. The only other bias values that are statistically significantly different from zero are for health insurance, household income and mother working. Table 14 shows the cross-validated bandwidth values that were used in generating the nonparametric matching estimates.

D. Performance Metrics

In table 15, we compute means of the absolute values of the percent difference and the difference in effect size units between the experimental and nonexperimental impact estimates. We group the models based on whether they rely on conditional independence with or without including lagged variables and whether the models are parametric regression or matching. In the cognitive domain, the estimators that include lagged values do well, and the difference-in-differences matching estimator, in particular, gets within 7% (0.14 δ) of the HSIS impact estimates. In the health domain, difference-in-differences matching also performs the best, although the bias is substantial for all the models. In the parenting domain and in the labor domain, the variation in performance across models is less pronounced. The largest biases

37 See Heckman and Hotz (1989) for a more detailed discussion of the test and Heckman et al. (1999) for a discussion of caveats regarding its use. See Ham, Li, and Reagan (2011) for an application of pre-program specification tests in the context of controlling for selectivity in estimating the returns to migration.
### Table 12
Heckman-Hotz Tests for Regression-Based Estimators

<table>
<thead>
<tr>
<th>ECLS-B 3-Year-Olds</th>
<th>Regression with Controls</th>
<th>Regression with Controls&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Difference-in-Differences with Controls</th>
<th>ECLS-B 4-Year-Olds</th>
<th>Lagged Value-Added with Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>−.51</td>
<td>(1.77)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics percentile</td>
<td>−1.18</td>
<td>(2.07)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child health good/excellent</td>
<td>.02</td>
<td>(0.02)</td>
<td>−.03</td>
<td>−.06</td>
<td>−.07</td>
</tr>
<tr>
<td>Health insurance</td>
<td>.02</td>
<td>(0.01)</td>
<td>.02</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>Parenting:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>.01</td>
<td>(0.04)</td>
<td>.07</td>
<td>.03</td>
<td>−.01</td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>.02</td>
<td>(0.04)</td>
<td>.04</td>
<td>.08</td>
<td>.04</td>
</tr>
<tr>
<td>Used time out</td>
<td>.03</td>
<td>(0.04)</td>
<td>.04</td>
<td>−.00</td>
<td>.03</td>
</tr>
<tr>
<td>Spanked child</td>
<td>−.00</td>
<td>(0.04)</td>
<td>−.03</td>
<td>.04</td>
<td>.04</td>
</tr>
<tr>
<td>Labor:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td>35</td>
<td>(45)</td>
<td>20</td>
<td>4,318</td>
<td>1,413</td>
</tr>
<tr>
<td>Mother works</td>
<td>−.04</td>
<td>(0.04)</td>
<td>−.10</td>
<td>−.14&lt;sup&gt;b&lt;/sup&gt;</td>
<td>−.12</td>
</tr>
<tr>
<td>Father works (if present)</td>
<td>.03</td>
<td>(0.03)</td>
<td>−.03</td>
<td>−.02</td>
<td>.08</td>
</tr>
</tbody>
</table>

*Note.—Standard errors are reported in parentheses.

<sup>a</sup> Control variables include all the variables listed in table 1.

<sup>b</sup> Statistically significant difference from zero at the 5% significance level.
are observed for the income outcome measure. As noted in Section IV.C, the survey instruments for income between the HSIS and ECLS-B displayed the largest qualitative differences. Also, HSIS and ECLS-B parents are located in different labor markets, which was shown to be an important determinant of performance in the Heckman, Ichimura, and Todd (1997) study.

In general, there is substantially more variance in performance metrics across outcome domains than across models, which can been seen comparing the overall means for the model along row with the overall means for the domains along the column. One possible explanation for such a result is the CIA and BSA assumptions that justify application of the estimators hold for some outcomes and not for others. The small biases observed for the cognitive outcomes both for the difference-in-differences matching and for the

### Table 13
**Heckman-Hotz Tests for Local Linear Matching Estimators**

<table>
<thead>
<tr>
<th></th>
<th>Cross-Section</th>
<th>Difference-in-Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ECLS-B 3-Year-Olds</td>
<td>ECLS-B 4-Year-Olds</td>
</tr>
<tr>
<td><strong>Cognitive:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>$-7.07^*$</td>
<td>(2.88)</td>
</tr>
<tr>
<td>Mathematics percentile</td>
<td>$-6.36^*$</td>
<td>(3.24)</td>
</tr>
<tr>
<td><strong>Health:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child health good/excellent</td>
<td>.03</td>
<td>-.00</td>
</tr>
<tr>
<td>Health insurance</td>
<td>.02*</td>
<td>.04*</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.05)</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td>(.02)</td>
</tr>
<tr>
<td><strong>Parenting:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>-.01</td>
<td>.07</td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>.01</td>
<td>.12</td>
</tr>
<tr>
<td>Used time out</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>Spanked child</td>
<td>-.01</td>
<td>-.06</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td>(.05)</td>
</tr>
<tr>
<td><strong>Labor:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td>$-3461^*$</td>
<td>-521</td>
</tr>
<tr>
<td></td>
<td>(1,477)</td>
<td>(1,296)</td>
</tr>
<tr>
<td>Mother works</td>
<td>-.06</td>
<td>-.17*</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td>(.06)</td>
</tr>
<tr>
<td>Father works (if present)</td>
<td>.02</td>
<td>-.01</td>
</tr>
<tr>
<td></td>
<td>(.03)</td>
<td>(.07)</td>
</tr>
</tbody>
</table>

**NOTE.**—ECLS-B 3-year-olds have no baseline reading or math scores. Difference-in-differences matching estimators can only be implemented for the 4-year-olds because the 3-year-olds do not have prebaseline outcomes to form the differenced outcome. Standard errors are reported in parentheses.  

* Statistically significant difference from zero at the 5% significance level.
regression models that include lagged outcomes are noteworthy given debates in the recent literature about the performance of value-added models for test scores (see e.g., Rothstein 2010; Chetty, Friedman, and Rockoff 2014). They support the use of value-added models applied to nonexperimental data.

In the bottom panel of table 15, we calculate the average bias for the subset of the estimators that passed the Heckman-Hotz tests. Because of missing baseline outcome data, only 51% of the outcomes and models were eligible for these tests, and 91% passed the Heckman-Hotz tests among those that were eligible, which suggests that these tests are not very discriminating in these data. Overall, the average bias in percent increases from 362 to 398 but slightly decreases in effect size from 1.62 to 1.60. This masks some heterogeneity across models (cross-section matching improves) and across domains (labor improves overall, but other domains become worse).

### E. Robustness

Finally, in table 16, we consider the performance of some alternative ways of implementing the estimators as a robustness check. First, for the regression models, we used inverse probability weights constructed using the propensity scores. Such a model is one example of a “doubly robust” estimator that can produce unbiased estimates if either the outcome equation or the

---

Table 14
Cross-Validated Bandwidths for Local Linear Matching

<table>
<thead>
<tr>
<th></th>
<th>Cross-Section</th>
<th>Difference-in-Differences</th>
<th>Heckman-Hotz</th>
<th>Cross-Section</th>
<th>Difference-in-Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3-Year-Olds</td>
<td>4-Year-Olds</td>
<td>3-Year-Olds</td>
<td>4-Year-Olds</td>
<td>3-Year-Olds</td>
</tr>
<tr>
<td>Cognitive:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading percentile</td>
<td>.05</td>
<td>.14</td>
<td>.05</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>Mathematics percentile</td>
<td>.40</td>
<td>.15</td>
<td>.07</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>Health:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child health</td>
<td>.09</td>
<td>.06</td>
<td>.40</td>
<td>.08</td>
<td>.02</td>
</tr>
<tr>
<td>Health insurance</td>
<td>.01</td>
<td>.07</td>
<td>.02</td>
<td>.34</td>
<td>.04</td>
</tr>
<tr>
<td>Parenting:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read to child</td>
<td>.08</td>
<td>.05</td>
<td>.34</td>
<td>.40</td>
<td>.04</td>
</tr>
<tr>
<td>Parental safety practices</td>
<td>.10</td>
<td>.25</td>
<td>.06</td>
<td>.40</td>
<td>.08</td>
</tr>
<tr>
<td>Used time out</td>
<td>.02</td>
<td>.07</td>
<td>.40</td>
<td>.11</td>
<td>.02</td>
</tr>
<tr>
<td>Spanked child</td>
<td>.10</td>
<td>.40</td>
<td>.16</td>
<td>.02</td>
<td>.40</td>
</tr>
<tr>
<td>Labor:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td>.02</td>
<td>.03</td>
<td>.02</td>
<td>.40</td>
<td>.02</td>
</tr>
<tr>
<td>Mother works</td>
<td>.40</td>
<td>.40</td>
<td>.10</td>
<td>.40</td>
<td>.40</td>
</tr>
<tr>
<td>Father works (if present)</td>
<td>.40</td>
<td>.12</td>
<td>.08</td>
<td>.05</td>
<td>.40</td>
</tr>
</tbody>
</table>

**Note.**—We conducted a grid search on [0.01,0.40] with a .01 step size.
propensity score model is correctly specified. Relative to the baseline regression models, inverse probability weighting actually increases the bias. This is consistent with some work arguing that these methods can be sensitive to small degrees of misspecification (Kang and Schafer 2007). Finally, we tried estimating propensity score models without interaction effects, as is more commonly done in the literature. The results again are not as good as our baseline models.

Table 15
Bias Summary Statistics by Model and Domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>Regression without Lagged Variables</th>
<th>Regression with Lagged Variables</th>
<th>Cross-Section Matching</th>
<th>Difference-in-Differences Matching</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>74</td>
<td>25</td>
<td>107</td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>(.82)</td>
<td>(.66)</td>
<td>(1.86)</td>
<td>(.14)</td>
<td>(.90)</td>
</tr>
<tr>
<td>Health</td>
<td>306</td>
<td>494</td>
<td>361</td>
<td>232</td>
<td>411</td>
</tr>
<tr>
<td></td>
<td>(.59)</td>
<td>(.57)</td>
<td>(.50)</td>
<td>(.42)</td>
<td>(.54)</td>
</tr>
<tr>
<td>Parenting</td>
<td>210</td>
<td>248</td>
<td>191</td>
<td>196</td>
<td>227</td>
</tr>
<tr>
<td></td>
<td>(1.73)</td>
<td>(2.15)</td>
<td>(1.76)</td>
<td>(2.01)</td>
<td>(2.01)</td>
</tr>
<tr>
<td>Labor (excluding income)</td>
<td>219</td>
<td>454</td>
<td>573</td>
<td>718</td>
<td>475</td>
</tr>
<tr>
<td></td>
<td>(1.22)</td>
<td>(1.48)</td>
<td>(1.20)</td>
<td>(1.51)</td>
<td>(1.41)</td>
</tr>
<tr>
<td>Income</td>
<td>699</td>
<td>1,023</td>
<td>1,241</td>
<td>793</td>
<td>975</td>
</tr>
<tr>
<td></td>
<td>(2.53)</td>
<td>(3.73)</td>
<td>(4.49)</td>
<td>(2.89)</td>
<td>(3.55)</td>
</tr>
<tr>
<td>Overall</td>
<td>249</td>
<td>394</td>
<td>372</td>
<td>349</td>
<td>362</td>
</tr>
<tr>
<td></td>
<td>(1.34)</td>
<td>(1.71)</td>
<td>(1.69)</td>
<td>(1.49)</td>
<td>(1.62)</td>
</tr>
</tbody>
</table>

Only models that passed Heckman-Hotz:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Regression without Lagged Variables</th>
<th>Regression with Lagged Variables</th>
<th>Cross-Section Matching</th>
<th>Difference-in-Differences Matching</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>49</td>
<td>49</td>
<td>(1.16)</td>
<td>(1.16)</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>306</td>
<td>814</td>
<td>580</td>
<td>377</td>
<td>603</td>
</tr>
<tr>
<td></td>
<td>(.59)</td>
<td>(.67)</td>
<td>(.44)</td>
<td>(.16)</td>
<td>(.56)</td>
</tr>
<tr>
<td>Parenting</td>
<td>210</td>
<td>313</td>
<td>191</td>
<td>252</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>(1.73)</td>
<td>(2.20)</td>
<td>(1.76)</td>
<td>(2.12)</td>
<td>(1.99)</td>
</tr>
<tr>
<td>Labor (excluding income)</td>
<td>219</td>
<td>263</td>
<td>141</td>
<td>1,086</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>(1.22)</td>
<td>(.95)</td>
<td>(1.15)</td>
<td>(.87)</td>
<td>(1.05)</td>
</tr>
<tr>
<td>Labor (excluding income)</td>
<td>699</td>
<td>937</td>
<td>818</td>
<td>690</td>
<td>832</td>
</tr>
<tr>
<td></td>
<td>(2.53)</td>
<td>(3.53)</td>
<td>(3.08)</td>
<td>(2.60)</td>
<td>(3.11)</td>
</tr>
<tr>
<td>Overall</td>
<td>264</td>
<td>495</td>
<td>281</td>
<td>514</td>
<td>398</td>
</tr>
<tr>
<td></td>
<td>(1.42)</td>
<td>(1.78)</td>
<td>(1.54)</td>
<td>(1.46)</td>
<td>(1.60)</td>
</tr>
</tbody>
</table>

**Note.**—Each cell contains the average of the absolute value of percentage difference from the HSIS impact (the top number) and the average of the absolute value of the difference from the HSIS impact in σ units (in parentheses). The average is across all models and outcomes within a cell. For models that passed Heckman-Hotz tests, an empty cell indicates that no models passed.

*This contains both difference-in-differences and lagged value-added models.

38 For recent work on other econometric estimators with the doubly robust property, see Klein (2011) and Graham et al. (2016).
Table 16
Bias Summary Statistics for Robustness Checks

<table>
<thead>
<tr>
<th></th>
<th>Regression without Lagged Variables</th>
<th>Regression with Lagged Variables</th>
<th>Cross-Section Matching</th>
<th>Difference-in-Differences Matching</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive</td>
<td>74.00</td>
<td>25.00</td>
<td>107.00</td>
<td>7.00</td>
<td>52.00</td>
</tr>
<tr>
<td></td>
<td>.82 .66</td>
<td>1.86 .14</td>
<td>.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>306.00</td>
<td>494.00</td>
<td>361.00</td>
<td>232.00</td>
<td>411.00</td>
</tr>
<tr>
<td></td>
<td>.59 .57</td>
<td>.50 .42</td>
<td>.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parenting</td>
<td>210.00</td>
<td>248.00</td>
<td>191.00</td>
<td>196.00</td>
<td>227.00</td>
</tr>
<tr>
<td></td>
<td>1.73 2.15</td>
<td>1.76 2.01</td>
<td>2.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labor (excluding income)</td>
<td>219.00</td>
<td>454.00</td>
<td>573.00</td>
<td>718.00</td>
<td>475.00</td>
</tr>
<tr>
<td></td>
<td>1.22 1.48</td>
<td>1.20 1.51</td>
<td>1.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>699.00</td>
<td>1,023.00</td>
<td>1,241.00</td>
<td>793.00</td>
<td>975.00</td>
</tr>
<tr>
<td></td>
<td>2.53 3.73</td>
<td>4.49 2.89</td>
<td>3.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>249.00</td>
<td>394.00</td>
<td>372.00</td>
<td>349.00</td>
<td>362.00</td>
</tr>
<tr>
<td></td>
<td>1.34 1.71</td>
<td>1.69 1.49</td>
<td>1.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inverse probability weightsb:</td>
<td>80.00</td>
<td>36.00</td>
<td>51.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive</td>
<td>1.00 .91</td>
<td>.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>385.00</td>
<td>443.00</td>
<td>431.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.76 .64</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parenting</td>
<td>253.00</td>
<td>338.00</td>
<td>321.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.83 2.39</td>
<td></td>
<td>2.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labor (excluding income)</td>
<td>552.00</td>
<td>411.00</td>
<td>439.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.52 1.45</td>
<td></td>
<td>1.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>824.00</td>
<td>1,040.00</td>
<td>997.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.00 3.80</td>
<td></td>
<td>3.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>352.00</td>
<td>414.00</td>
<td>400.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.54 1.85</td>
<td></td>
<td>1.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main effects only:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive</td>
<td>82.00</td>
<td>37.00</td>
<td>154.00</td>
<td>27.00</td>
<td>72.00</td>
</tr>
<tr>
<td></td>
<td>1.08 1.01</td>
<td>2.24 .72</td>
<td>1.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>191.00</td>
<td>391.00</td>
<td>305.00</td>
<td>723.00</td>
<td>398.00</td>
</tr>
<tr>
<td></td>
<td>.31 .42</td>
<td>.27 .66</td>
<td>.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parenting</td>
<td>142.00</td>
<td>239.00</td>
<td>154.00</td>
<td>197.00</td>
<td>207.00</td>
</tr>
<tr>
<td></td>
<td>1.61 2.12</td>
<td>1.68 1.87</td>
<td>1.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labor (excluding income)</td>
<td>361.00</td>
<td>496.00</td>
<td>673.00</td>
<td>602.00</td>
<td>517.00</td>
</tr>
<tr>
<td></td>
<td>1.16 1.43</td>
<td>1.20 1.53</td>
<td>1.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>938.00</td>
<td>1,168.00</td>
<td>1,115.00</td>
<td>1,211.00</td>
<td>1,134.00</td>
</tr>
<tr>
<td></td>
<td>3.41 4.27</td>
<td>4.05 4.43</td>
<td>4.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>252.00</td>
<td>394.00</td>
<td>363.00</td>
<td>467.00</td>
<td>378.00</td>
</tr>
<tr>
<td></td>
<td>1.36 1.74</td>
<td>1.65 1.70</td>
<td>1.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Each cell contains the average of the absolute value of percentage difference from the HSIS impact (the top number) and the average of the absolute value of the difference from the HSIS impact in $\sigma$ units (the bottom number in parentheses). The average is across all models and outcomes within a cell.

a This contains both difference-in-differences and lagged value-added models.

b Inverse probability weights are only applicable to regression models.
VI. Summary

This paper studies the effectiveness of nonexperimental methods applied to data from the ECLS-B for the purpose of evaluating the effects of the Head Start Program. It compares experimental estimates derived from the HSIS data to nonexperimental estimates obtained from ECLS-B for four different outcome domains: cognitive test scores, child health, parenting behaviors, and parent labor market outcomes. The nonexperimental estimators include parametric cross-section regression and difference-in-differences regression as well as nonparametric cross-section matching and difference-in-differences matching. Pre-program exogeneity tests were not found to be very discriminating in isolating the best performing estimators/outcomes. Overall, we find that the difference-in-differences matching estimator exhibits the best performance in terms of having the lowest bias values and capturing the pattern of statistically significant treatment effects across different outcome domains (as shown in table 11). The estimator comes remarkably close to replicating the experimental impact estimates for both reading and math cognitive scores. We find that the lowest bias estimates are obtained with a propensity score specification that includes interaction terms to achieve better balance in the covariates. The difference-in-differences and value-added regression models also exhibit relatively low bias values, indicating the importance of using an estimator that incorporates information on lagged outcome variables.

Another finding that emerged from our study is that the estimated bias varies substantially across outcome measures, even more so than across methods. Outcomes, such as child test scores, tend to have smaller biases, regardless of the estimation method, but other outcomes, such as household income, exhibit consistently large biases. This difference is likely due to differences in how income is measured across survey instruments, as well as possibly to different income levels across local labor markets.\textsuperscript{39} The literature has focused a lot on the question of which estimation methods are reliable, but in our study, reliability depends on the outcome to which the estimator is applied.

References


\textsuperscript{39} Questionnaire differences and local labor market differences were shown to be important in Heckman, Ichimura, and Todd (1997).


Black, Dan, Jose Galdo, and Jeffrey A. Smith. 2007. Evaluating the bias of the regression discontinuity design using experimental data. Working paper, University of Chicago.


Griffen, Andrew S. 2016. Evaluating the effects of child care policies on children’s cognitive development and maternal labor supply. Unpublished manuscript, University of Tokyo.


Heckman, James, Neil Hohmann, and Jeffrey Smith, with Michael Khoo. 2000. Substitution and dropout bias in social experiments: A study of an


Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. Characterizing selection bias using experimental data. *Econometrica* 66, no. 5:1017–98. (Note: Reference citations in the text to Heckman et al. [1998] refer to this reference.)


Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, Michael Lopez, Nicholas Zill, Gary and Shapiro, Pam Broene, Debra Mekos, Monica Rohacek, et al. 2005. Head Start Impact Study: First year findings. Ad-
ministration for Children and Families, Department of Health and Human Services, Washington, DC.
White House. 2011. We can’t wait: President Obama takes action to improve quality and promote accountability in Head Start programs. Office of the Press Secretary.


